AGU100 ADVANCING EARTH AND SPACE SCIENCE

Water Resources Research

RESEARCH ARTICLE
10.1029/2017WR021895

Key Points:
- 1023 combinations of 9 remotely sensed products were explored for constraining parameters in the absence of streamflow data
- Combining multiple (satellite) data sets for deriving posterior parameter ranges leads to a narrower parameter search space
- Especially the soil moisture products of AMSR-E, ASCAT, and the total water storage anomalies from GRACE helped in determining feasible parameter sets with good performance in streamflow prediction

Supporting Information:
- Supporting Information S1

Correspondence to:
R. C. Nijzink,
r.c.nijzink@tudelft.nl

Citation:
Nijzink, R. C., Almeida, S., Pechlivanidis, I. G., Capell, R., Gustafssons, D., Arheimer, B., et al. (2018). Constraining conceptual hydrological models with multiple information sources. *Water Resources Research*, *54*. https://doi.org/10.1029/2017WR021895

Received 19 SEP 2017
Accepted 17 SEP 2018
Accepted article online 24 SEP 2018

Constraining Conceptual Hydrological Models With Multiple Information Sources

R. C. Nijzink[1,2], S. Almeida[3,4], I. G. Pechlivanidis[5], R. Capell[5], D. Gustafssons[5], B. Arheimer[5], J. Parajka[6], J. Freer[7,8], D. Han[3,8], T. Wagener[3,8], R. R. P. van Nooijen[1], H. H. G. Savenije[1], and M. Hrachowitz[1]

[1]Water Resources Section, Faculty of Civil Engineering and Geosciences, Delft University of Technology, Delft, Netherlands, [2]Now at Catchment and Eco-Hydrology Research Group, Environmental Research and Innovation Department, Luxembourg Institute of Science and Technology, Belvaux, Luxembourg, [3]Department of Civil Engineering, University of Bristol, Bristol, UK, [4]Now at School of Civil Engineering, University of Leeds, Leeds, UK, [5]Swedish Meteorological and Hydrological Institute, Norrköping, Sweden, [6]Institute of Hydraulic Engineering and Water Resources Management, Vienna University of Technology, Vienna, Austria, [7]School of Geographical Sciences, University of Bristol, Bristol, UK, [8]Cabot Institute, University of Bristol, Bristol, UK

Abstract The calibration of hydrological models without streamflow observations is problematic, and the simultaneous, combined use of remotely sensed products for this purpose has not been exhaustively tested thus far. Our hypothesis is that the combined use of products can (1) reduce the parameter search space and (2) improve the representation of internal model dynamics and hydrological signatures. Five different conceptual hydrological models were applied to 27 catchments across Europe. A parameter selection process, similar to a likelihood weighting procedure, was applied for 1,023 possible combinations of 10 different data sources, ranging from using 1 to all 10 of these products. Distances between the two empirical distributions of model performance metrics *with* and *without* using a specific product were determined to assess the added value of a specific product. In a similar way, the performance of the models to reproduce 27 hydrological signatures was evaluated relative to the unconstrained model. Significant reductions in the parameter space were obtained when combinations included Advanced Microwave Scanning Radiometer - Earth Observing System and Advanced Scatterometer soil moisture, Gravity Recovery and Climate Experiment total water storage anomalies, and, in snow-dominated catchments, the Moderate Resolution Imaging Spectroradiometer snow cover products. The evaporation products of Land Surface Analysis - Satellite Application Facility and MOD16 were less effective for deriving meaningful, well-constrained posterior parameter distributions. The hydrological signature analysis indicated that most models profited from constraining with an increasing number of data sources. Concluding, constraining models with multiple data sources simultaneously was shown to be valuable for at least four of the five hydrological models to determine model parameters in absence of streamflow.

1. Introduction

Computational techniques have been advancing and an abundance of new sources of information has become available over the recent years, but selecting meaningful parameters for catchment-scale hydrological models, in particular for predictions in ungauged catchments, remains problematic (Blöschl et al., 2013; Hrachowitz et al., 2013; Sivapalan, 2003) and is further exacerbated by the worldwide ongoing reductions of stream gauging networks (Fekete & Vörösmarty, 2002; Hannah et al., 2011; Sivapalan, 2003).

The dependency on streamflow data for model calibration can, to a certain extent, be reduced by directly estimating individual model parameters (or at least defining nonuniform parameter prior distributions) from exploiting their links with readily available observations of other quantities than streamflow, which are observable at the scale of the model application, such as topographic considerations (Smith et al., 2016) or the long-term water balance (e.g., de Boer-Euser et al., 2016; Gao, Hrachowitz, Schymanski, et al., 2014; Nijzink et al., 2016). Similarly, when no streamflow observations are available, traditional regionalization techniques use climatic and physiographic data, to establish transfer functions that allow an indirect estimation of the actual model parameters (Götzinger & Bárdossy, 2007; Hundecha et al., 2016; Hundecha & Bárdossy, 2004; Merz & Blöschl, 2004; Samaniego et al., 2010; Wagener & Wheater, 2006). Alternatively, catchment

signatures can be used to condition the parameter space, as an alternative to streamflow calibration (Almeida et al., 2016; Bárdossy, 2007; Bulygina et al., 2009; Castiglioni et al., 2010, 2011; Yadav et al., 2007). In addition, model constraints or limits of acceptability (Beven, 2006), based on qualitative information without clear target values, often referred to as *soft data* (Seibert & McDonnell, 2002; Winsemius et al., 2009) or *expert knowledge* (Gharari et al., 2014; Kelleher et al., 2017; Pechlivanidis & Arheimer, 2015), meant to avoid physically implausible representations of the system, were in the past shown to be valuable to limit the feasible model parameter space (Freer et al., 2004; Hrachowitz et al., 2014). These constraints can be implemented either as a priori defined inequality constraints on parameters or on processes (Ambroise et al., 1996). The latter allows to contain the dynamics of individual model components to some degree (cf. Gharari et al., 2014; Wagener & Montanari, 2011), such as limiting long-term evaporation to values expected from the Budyko curve (e.g., Gerrits et al., 2009).

The increasing availability of remotely sensed data may provide ample opportunities to further constrain hydrological models and their parameters. While several recent reviews highlight their potential for applications in hydrology (e.g., AghaKouchak et al., 2015; Hrachowitz & Clark, 2017; Pechlivanidis & Arheimer, 2015; Xu et al., 2014), it can also be argued that remotely sensed high-resolution streamflow data are rather far from becoming a reality (Lettenmaier et al., 2015). Although successful attempts of using remotely sensed streamflow for model calibration have been reported (e.g., Sun et al., 2015; Tourian et al., 2017), the specific orbits of the observation satellites lead to spatial and temporal limitations, and only larger rivers can be monitored due to the large resolution. In contrast, products providing estimates of evaporation have in the past been shown to have considerable value for model applications, as summarized by several studies that point at the different advantages and disadvantages of these products (e.g., Verstraeten et al., 2008; Zhang et al., 2016). Besides evaporation products, the central importance of soil moisture and snow storage for the Earth's water cycle made it a focus of research efforts in the remote sensing community, which developed several satellite missions dedicated to soil moisture and snow cover mapping, such as the Soil Moisture and Ocean Salinity (SMOS; Kerr et al., 2012), Soil Moisture Active and Passive (SMAP; Brown et al. 2013), or NASA's Earth Observing System (Greenstone & King, 1999) missions. Furthermore, the Gravity Recovery and Climate Experiment (GRACE; Tapley et al., 2004) led to new, valuable information on total water storage based on remotely sensed gravity anomalies. These are just a few examples, while more remotely sensed products are currently available and new satellite missions are planned (e.g., GRACE-FO and SWOT), which will further increase the information available for hydrological modeling.

The challenge remains, though, how to select data that are suitable for use in hydrological model applications and to assess how they can support the modeling process in a meaningful and effective way. So far, information from remote sensing has been incorporated in applications of hydrological models in several ways. For example, data assimilation techniques are commonly used to update the states of a model (e.g., Liu et al., 2012; Liu & Gupta, 2007; Reichle, 2008). This can help to improve internal model dynamics and the resulting hydrological predictions (Crow & Ryu, 2009; Tangdamrongsub et al., 2015). Yet, it can be argued that the added value of data assimilation is actually an indicator of inadequate model parameters and/or model formulations (Spaaks & Bouten, 2013). Alternatively and directly addressing this issue, remotely sensed data can be directly used as calibration variables and thus to select feasible model parameters (e.g., Immerzeel & Droogers, 2008; Lopez Lopez et al., 2017; Pechlivanidis & Arheimer, 2015; Sutanudjaja et al., 2014). Although the above strategies are in principle a valid way forward, spatial and temporal mismatches between hydrological models and remotely sensed data (Vereecken et al., 2008; Xu et al., 2014) place some limitations on the value of these data. Acknowledging this, several new techniques are reported in the literature with the focus on, for example, the spatial patterns of remotely sensed data (e.g., Demirel et al., 2018; Githui et al., 2015; Stisen et al., 2008). It can be argued that especially the development of a pattern based objective function (e.g., Zink et al., 2018) is needed to optimally use the distributed information of the products. Correctly relating the spatial patterns to the models also mitigates the fact that hydrological variables are not directly observed by most remote sensors but rather inferred from models that link the observed variable with some hydrologically relevant variable, thus introducing an additional source of uncertainty. As a result there is a shift from using the absolute numbers obtained by remote sensing products to using those numbers more relatively, with the spatial patterns as an example.

A large number of studies previously assessed the added value of different remote sensing products, either for data assimilation or model calibration. These studies generally focused either on a single remote sensing

product, for example, GRACE (e.g., Lo et al., 2010; Mulder et al., 2015; Rakovec, Kumar, Attinger, et al., 2016; Werth et al., 2009), SCAT soil moisture (e.g., Parajka et al., 2009), and Advanced SCATterometer (ASCAT) soil moisture (Brocca et al., 2010), or on one-single model state or flux with a combination of products such as soil moisture (Wanders et al., 2014), which was all done with different levels of success. For example, Rakovec, Kumar, Mai, et al. (2016) showed that the addition of GRACE improved internal states of the model, but remotely sensed soil moisture deteriorated model performance. Nevertheless, the combined effects of several products, which deal with multiple model states and fluxes simultaneously, have only recently gained some attention, but this remains rather limited to two different model states or fluxes (Kunnath-Poovakka et al., 2016; Lopez Lopez et al., 2017; Tian et al., 2017). Full bootstrap procedures where multiple combinations of remote sensing products are tested have not been reported so far.

Thus, the objective of this paper is to explore the value of combining several types of remotely sensed data products that reflect different water balance components, to effectively and consistently constrain the parameter space of five different lumped conceptual hydrological models, as a stepping stone toward using combined remote sensing products also in more distributed modeling approaches. Even though the distributed nature of the products is not used up to its full potential by using lumped models, in this way we test the hypotheses that the combined use of different remote sensing products can (1) identify unfeasible parameter sets and thus reduce the feasible parameter space in order to shift toward higher average model performances and (2) improve the representation of model internal dynamics and hydrological signatures in comparison with a range of benchmarking streamflow performances.

2. Methodology

A detailed stepwise description of this experiment, with the model codes and links to the data, can be found in an online experiment protocol (http://dl-ng005.xtr.deltares.nl/view/66/) as part of the Virtual Water Science Laboratory of the SWITCH-ON project (Sharing Water-Related Information to Tackle Changes in the Hydrosphere - for Operational Needs). This protocol is developed to facilitate full experiment reproducibility and repeatability, according to the requirements suggested by Ceola et al. (2015).

2.1. Study Areas

A set of 27 European catchments was selected in order to cover a variety of landscapes, climates, and vegetation. The study sites included lowland catchments in the UK and Germany as well as more mountainous catchments in Austria and France. The study catchments also exhibit considerable climatic differences with aridity indices ranging from 0.5 to 1.1 (mean potential evaporation divided over the mean precipitation) and mean areal precipitation from 627 to 1593 mm/year. The selection was based on these differences, as well as on the length of the available time series (approximately 10 years) for the recent years in order to be comparable to the remotely sensed products.

Estimates of daily potential evaporation were derived from ERA-Interim data (air temperature, dew point temperature, wind speed, longwave radiation, and shortwave radiation) according to the Penman formulation as prescribed by FAO (Allen et al., 1998), and the air temperature of the ERA-Interim data (Dee et al., 2011) was also applied as forcing data for the snow modeling. Daily precipitation was derived from the Multi-Source Weighted-Ensemble Precipitation data set (MSWEP; Beck et al. 2017). Time series of streamflow covering recent years and with sufficient length (approximately 10 years of data) were for most catchments obtained from the Global Runoff Data Centre. In addition, three catchments were selected from the Hydrographic Service of Austria, and three from, respectively, the Hydrographic Service of the Autonomous Province of Bolzano, Regional Agency for the Protection of the Environment - Piedmont Region, and Regional Hydrologic Service - Tuscany Region. An overview of the catchments is provided in Table 1 and Figure 1.

2.2. Models

Five different rainfall-runoff models were applied to account for different model structures, which are briefly described here. These models were selected as these are widely used across Europe and differed conceptually. For more details about model structures, parameters, and prior parameter ranges the reader is referred to the supporting information S1.

Table 1
Overview of the Study Catchments, Their Characteristics, and the Modeled Time Series

	River	GRDC number	Area (km^2)	Aridity (−)	Prec. (mm/year)	Q (mm/year)	Warm start	Warm end	Cal start	Cal end
1	Gadera	-	394.0	0.52	1,095	670	10-1-1998	30-9-1999	10-1-1999	30-9-2009
2	Tanaro	-	500.0	0.81	1,059	625	10-1-2002	30-9-2003	10-1-2003	30-9-2012
3	Arno	-	751.0	0.81	1,069	398	10-1-2002	30-9-2003	10-1-2003	30-9-2013
4	Vils	-	198.0	0.32	1,593	1338	10-1-1999	30-9-2000	10-1-2000	30-9-2010
5	Grossarler	-	144.0	0.36	1,508	1202	10-1-1999	30-9-2000	10-1-2000	30-9-2010
6	Große Mühl	-	253.0	0.55	1,156	739	10-1-1999	30-9-2000	10-1-2000	30-9-2010
7	Broye	6935390	396.0	0.50	1,219	610	10-1-1998	30-9-1999	10-1-1999	30-9-2009
8	Treene	6338800	481.0	0.71	917	435	10-1-1998	30-9-1999	10-1-1999	30-9-2004
9	Risle	6118165	146.8	1.00	751	304	10-1-2000	30-9-2001	10-1-2001	30-9-2011
10	Leyre	6119200	1586.9	0.98	913	283	10-1-2000	30-9-2001	10-1-2001	30-9-2011
11	Erdre	6123170	462.5	0.98	789	181	10-1-2000	30-9-2001	10-1-2001	30-9-2011
12	Layon	6123180	927.9	1.10	694	105	10-1-2005	30-9-2006	10-1-2006	30-9-2011
13	Glane	6123420	296.7	0.80	981	401	10-1-2000	30-9-2001	10-1-2001	30-9-2011
14	Dragne	6123700	116.6	0.75	996	435	10-1-2001	30-9-2002	10-1-2002	30-9-2011
15	Roubion	6139220	190.4	0.87	920	276	10-1-2000	30-9-2001	10-1-2001	30-9-2011
16	Azergues	6139360	333.1	0.84	887	317	10-1-2000	30-9-2001	10-1-2001	30-9-2011
17	Enning-Dalsaelve	6229100	633.8	0.57	978	646	10-1-2004	30-9-2005	10-1-2005	30-9-2014
18	Fyllean	6233150	263.4	0.62	937	749	10-1-2003	30-9-2004	10-1-2004	30-9-2014
19	Kinzig	6335125	955.0	0.49	1,344	744	10-1-2001	30-9-2002	10-1-2002	30-9-2012
20	Modau	6335165	90.6	0.99	705	234	10-1-2001	30-9-2002	10-1-2002	30-9-2012
21	Rodach	6335540	716.0	0.76	830	458	10-1-2001	30-9-2002	10-1-2002	30-9-2012
22	Pfinz	6335640	232.2	0.74	920	248	10-1-2002	30-9-2003	10-1-2003	30-9-2013
23	Hunte	6337050	1408.5	0.80	806	219	10-1-2001	30-9-2002	10-1-2002	30-9-2012
24	Wuemme	6337060	934.4	0.77	855	342	10-1-2001	30-9-2002	10-1-2002	30-9-2012
25	Deveron	6604850	954.6	0.49	1,030	636	10-1-2001	30-9-2002	10-1-2002	30-9-2012
26	Little Ouse	6606250	757.4	1.05	627	143	10-1-2003	30-9-2004	10-1-2004	30-9-2012
27	Stour	6606850	656.7	1.03	632	153	10-1-2001	30-9-2002	10-1-2002	30-9-2012

Note. GDRC = Global Runoff Data Centre. Dates are formatted as day/month/year.

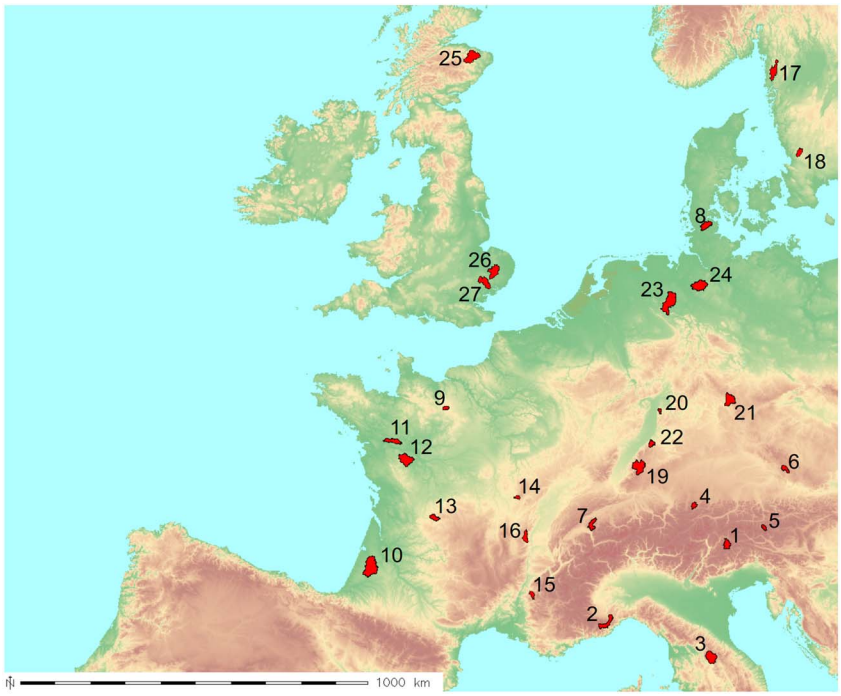

Figure 1. The 27 study catchments and their location in Europe. See Table 1 for the catchment characteristics.

2.2.1. FLEX

The FLEX model (Fenicia et al., 2008) is a lumped model that consists of four storage components and a snow module. The snow module, based on a degree-day approach, runs first and determines the effective precipitation consisting of rainfall and snowmelt. After this, the water enters an interception reservoir, from which intercepted water can evaporate and/or leave the reservoir after exceeding a certain threshold. The remaining precipitation after interception is split into runoff and infiltration in the subsequent step. The infiltrated water is stored in the soil moisture reservoir, from which transpiration takes place. A portion of the runoff goes to a fast reservoir, another portion to the groundwater reservoir through preferential percolation. The model uses eight parameters that are left free for calibration.

2.2.2. FLEXtopo

The FLEXtopo model (Savenije, 2010) uses hydrological response units based on different landscape elements to capture the core processes for different parts in the landscape. In this setup, the landscape units were defined as plateau, hillslope, and wetland, similar to previous applications (de Boer-Euser et al., 2017; Gao, Hrachowitz, Fenicia, et al., 2014; Gharari et al., 2014). For each model unit, a snow routine is followed by an interception reservoir and unsaturated reservoir. For plateau landscapes, recharge to the groundwater can happen through matrix percolation, as a function of soil moisture, and preferential percolation through macropores or cracks and fissures. In contrast, the hillslope areas are only assumed to contribute to the groundwater through preferential percolation and the wetlands even receive water from the groundwater reservoir through capillary rise. In the original application (Gao, Hrachowitz, Fenicia, et al., 2014) FLEXtopo uses proportionalities between parameters of different landscape classes (e.g., interception capacity of forest bigger than grass), which limit the feasible parameter space. Here in this comparative analytical framework additional conditions were not implemented for FLEXtopo, leaving a relatively wide parameter space. In total, 24 parameters are left free for calibration.

2.2.3. HYMOD

The HYMOD model (Boyle, 2001; Wagener et al., 2001) runs first a snow module from which rainfall and snowmelt continue toward the unsaturated zone. Here evaporation is determined as a function of soil moisture and runoff is generated, based on the spatial distribution of maximum storage capacities in the catchment as defined by a reflected power function. This runoff is divided over a series of fast reservoirs and one slow reservoir. The contributions of the fast flows and slow flows eventually determine the final streamflow. In total, eight parameters are free for calibration.

2.2.4. HYPE

The HYPE model (Lindström et al., 2010) runs first a snow module, after which the model structure contains three soil layers with assigned soil depths. Water can evaporate from the first two layers, and runoff is generated when the maximum storage capacity of these layers is reached or when maximum infiltration capacities are exceeded. Water can percolate downward through matrix flow or preferentially through fast flow paths. The lowest soil layer reflects the groundwater contribution to the streamflow, and an additional aquifer routine can be applied. Eventually, a routing function is applied to the total outflows to obtain the final streamflow. A set of 22 model parameters was selected for optimization in the parameter selection procedures. The model setup applied here is based on the E-HYPE modeling setup (Donnelly et al., 2009), therefore catchments 6, 9, and 14 were not considered for this analysis as these are not part of the E-HYPE model.

2.2.5. TUW

The TUW model (Parajka et al., 2007) uses a similar model structure as originally applied in the HBV model (Bergström, 1992). First, a snow routine is run based on a degree-day approach, after which water enters the soil moisture reservoir and becomes available for evaporation. Here evaporation is determined as a function of soil moisture, and runoff is generated based on a function defining the spatial distribution of maximum storage capacities as well. Next, the water moves to a fast reservoir, which has an additional overflow outlet to represent a very fast component. Percolation from the fast reservoir toward the slow, groundwater reservoir takes place subsequently. In a last step, the sums of slow and fast runoff components are routed through the system with triangular lag functions. The TUW model has 15 parameters free for calibration.

2.3. Data Sources for Constraining Parameters

Nine different remote sensing products and an analytical framework, from four functionally similar groups, were tested in this study for their information content to select meaningful model parameters and thus to constrain the feasible parameter space. Each group provides information about a different component of

Table 2
Details of the Remote Sensing Products

Product	Version	Spatial resolution	Temporal resolution	Reference	Model state/flux	Performance metric
AMSR-E -LPRM	V2	25 × 25 km	daily	Owe et al. (2008)	Soil moisture	Squared correlation coefficient
ASCAT -SWI	V3	0.1°	daily	Wagner et al. (1999)	Soil moisture	Squared correlation coefficient
SMOS	V620	~ 15 km	2–3 days	Kerr et al. (2012)	Soil moisture	Squared correlation coefficient
NDII	V6	500 m	daily	Sriwongsitanon et al. (2016)	Soil moisture	Squared correlation coefficient
Budyko				Budyko (1974)		Relative error
LSA-SAF		3 km	daily	Ghilain et al. (2011)	Evaporation	Squared correlation coefficient
MOD16	V5	500 m	8-day	Mu et al. (2011)	Evaporation	Squared correlation coefficient
GRACE		1 °	30 days	Tapley et al. (2004)	Total water storage	Squared correlation coefficient
MOD10	V5	500 m	daily	Hall et al. (2006a)	Snow state	Squared correlation coefficient
MYD10	V5	500 m	daily	Hall et al. (2006b)	Snow state	Squared correlation coefficient

Note. AMSR-E = Advanced Microwave Scanning Radiometer - Earth Observing System; LPRM = Land Parameter Retrieval Model; ASCAT = Advanced SCATterometer; SWI = Soil Water Index; NDII = Normalized Difference Infrared Index; LSA-SAF = Land Surface Analysis - Satellite Application Facility; GRACE = Gravity Recovery and Climate Experiment; SMOS = Soil Moisture and Ocean Salinity.

the hydrological system: (1) soil moisture, (2) evaporation, (3) total water storage, and (4) snow accumulation. A general overview of the used products and the main specifications on, for example, spatial and temporal resolution can be found in Table 2.

The first group contains soil moisture estimates from four different remote sensing products. One of the soil moisture products used in this study is derived from the ASCAT on board the Metop satellite, which uses C band (5.255 GHz) to estimate surface soil moisture. Scatterometer data processed with the algorithm provided by Wagner et al. (1999) was used in this experiment, representing the Soil Water Index or the relative soil moisture in the root zone. The second soil moisture product comes from the Land Parameter Retrieval Model (Owe et al., 2008) with data from the Advanced Microwave Scanning Radiometer - Earth Observing System (AMSR-E, with C band and X band) as input and represents the top 2–3 cm of soil moisture. The last soil moisture product explored in this study is obtained from the SMOS (Kerr et al., 2012) mission, also representing the soil moisture in the upper centimeters of the soil (L band, 1–2 GHz). In addition, the Normalized Difference Infrared Index (NDII) was calculated based on MODerate Resolution Imaging Spectroradiometer (MODIS) images, as recent results suggest a link to root zone soil moisture storage (Sriwongsitanon et al., 2016). Even though most of the products represent only soil moisture in the top soil, the products were directly compared to the soil moisture states of the models, without adjustments or exclusions of specific days (e.g., excluding snow days). Therefore, it was assumed that at least a linear relationship exists between modeled soil moisture state and the observations of the soil moisture products, even though these do not represent exactly the same soil moisture state as the model. Thus, the squared correlation coefficient was used as a performance metric for all soil moisture products.

The second group contains evaporation estimates from two remote sensing products and the Budyko framework. Specifically, the daily product of Land Surface Analysis - Satellite Application Facility (LSA-SAF), as European Organisation for the Exploitation of Meteorological Satellites (EUMETSAT), was selected as well as the MOD16 product from the MODIS (Mu et al., 2011). The MOD16 8-day evaporation product is based on a Penman-Monteith approach, and the final product consists of soil evaporation, transpiration, and interception evaporation from the canopy. The LSA-SAF evaporation product uses a similar Penman-Monteith approach, but the products differ in, among other things, formulations of aerodynamic and stomatal resistances, the (absence of) explicit accounting for interception evaporation, the temporal resolution of the satellite, and the differences in other technical satellite specifications. Besides these remotely sensed daily and eight-daily products, the analytical Budyko framework (Budyko, 1974) was applied to obtain an additional long-term estimate of evaporation as model constraint. Also for this group of products, the squared correlation coefficient was used as a performance metric, only for the Budyko framework the relative error was used as this only comprises a single value instead of a time series. For the MOD16 8-day evaporation product the comparison was made for the modeled eight-day evaporation as well.

The third group provides estimates of changes in total water storage from one remote sensing product. To do that, data on gravity anomalies from the Gravity Recovery and Climate Experiment (GRACE; Tapley et al., 2004) were linked to water storage fluctuations. Similar to (Rakovec, Kumar, Attinger, et al., 2016), the

GRACE data from the three processing centers of Center for Space Research, University of Texas, USA, Geoforschungs Zentrum Potsdam, Germany, and Jet Propulsion Laboratory, USA, were merged into one combined product for the analysis. These products were originally corrected for atmospheric pressure and mass changes (Landerer & Swenson, 2012) and thus represent only total water storage anomalies. The GRACE water storage anomalies relative to the baseline of 2004–2009 were also corrected to represent the anomalies over the time series under consideration, which differed per study catchment. Even though the resolution of GRACE is relatively coarse compared to the catchments under consideration, we still hypothesize that this signal may be of help in relatively homogeneous areas and/or by the seasonality of the signal. The squared correlation coefficient between modeled total water storage anomalies and GRACE storage anomalies was used as performance metric again.

The fourth group contains information on snow accumulation and depletion pattern from two remote sensing products. The MODIS Terra Snow Cover (Hall et al., 2006a) and MODIS Aqua Snow Cover (Hall et al., 2006b) products both provide fractional snow cover on a daily basis and 500-m resolution. This snow cover is determined by the Normalized Difference Snow Index and relies on the fact that while clouds have large reflectance in both visible and infrared bands, snow has a larger reflectance only in the visible domain. The average was taken over all noncloud covered cells over a catchment, but this was only done when the cloud coverage did not exceed 60% of the catchment, similar to Parajka & Blöschl (2008), in order to determine reliable snow coverage values. These catchment-averaged snow coverage values were compared with modeled snow water equivalents (FLEX, FLEXtopo, and TUW), assuming a linear relation between coverage and snow water equivalent. Only for HYPE, comparing modeled and observed snow coverage was directly possible due to the more extensive snow module. The squared correlation coefficient was again computed as performance metric.

In general, all products were processed in order to be compatible with the model scales, which was in most cases the catchment scale, as the models were applied in a lumped manner. Thus, the average of the cells covering the catchment was determined and used in the analysis. Only for FLEXtopo, the products were averaged over a subarea of the catchment, the landscape units, whereas the other models all used catchment-averaged values. For example, the LSA-SAF evaporation was averaged over the catchment area defined as plateau landscape, in order to compare with the modeled evaporation from the plateau model structure.

2.4. Identifiability Analysis

An identifiability analysis was carried out in order to assess which products can be related to which parameters in a meaningful way. Hence, the results of this analysis will be used to avoid that parameters are constrained with products they are insensitive to. It is assured in this way that all results of constraining models with remotely sensed products can be related to the remotely sensed products itself.

In a first step, random parameter sets were generated for each of the five models by using Latin Hypercube sampling to achieve a somewhat homogenous exploration of the respective parameter spaces. The parameters were sampled from uniform prior distributions with parameter ranges set as wide as possible without becoming physically implausible. FLEX and HYMOD were sampled 80,000 times, whereas FLEXtopo, HYPE, and TUW were sampled 100,000 times due to the larger number of free calibration parameters. In other words, all free calibration parameters were generated simultaneously by Latin Hypercube sampling. These parameter sets were then used together with the daily input data to generate either 80,000 or 100,000 model realizations per model for each catchment, covering a time period of approximately 10 years (see Table 1).

Subsequently, it was evaluated how well the modeled state and/or flux variables of each model realization were able to reproduce the different data from the group of remote sensing products that correspond to that specific model state or flux. For example, modeled evaporation for each of the 80,000(100,000) model realizations for each catchment was evaluated against the different evaporation estimates provided by the data from group 2 (see section 2.3). The squared correlation coefficient was used as a performance metric for model evaluation against each remote sensing product, emphasizing the models' ability to reproduce the temporal dynamics of a given variable but ignoring the magnitude of the variable itself. For evaluation against long-term evaporation from the Budyko curve the relative error was used. See also Table 2 for which performance metric was used per product.

Based on these samples, an identifiability analysis was employed similar to Regional Sensitivity Analysis (Hornberger & Spear, 1980; Wagener & Kollat, 2007), which has in the past been widely used as a measure of sensitivity (Demaria et al., 2007; McIntyre et al., 2003; Sieber & Uhlenbrook, 2005). In this type of analysis, the maximum distance between the prior cumulative parameter distribution and the posterior cumulative distribution serves as an informal indication of sensitivity. Here the posterior parameter distributions were determined based on a weighting procedure (e.g., Freer et al., 1996):

$$L_2(\theta) = L(\theta)^n * L_0(\theta)/C \tag{1}$$

where L_0 is the prior probability density function (−), L_2 is the posterior probability density function (−), n is a weighting factor (set to 10; [−]), C is a normalizing constant (−), the parameter set θ consists of all model parameters, and $L(\theta)$ is an informal likelihood weight that is here the squared correlation coefficient for the remotely sensed products and the relative error for the Budyko framework.

2.5. Parameter Selection—Constraining Models Using Remote Sensing Data

The identifiability analysis was combined with simple reasoning (e.g., snow performance metrics should relate to snow parameters) to relate parameters to relevant performance metrics. Thus, the final weighting was carried out on the basis of one or a combination of m selected performance metrics to model states/fluxes, which the identifiability analysis suggests to be relevant for the parameter under consideration, and is hence able to identify the parameter without using observed discharge data. In this way, only sensitive parameters are constrained, assuring that the final results can be directly attributed in a meaningful way to the products used for the constraints. The same parameter sets obtained by Latin-Hypercube sampling for the identifiability analysis were used here. Table 3 gives an overview of the parameters and the remotely sensed data sources that these parameters were eventually linked to. For example, the new parameter ranges linked to snow processes were calculated using weights derived from the model's ability to reproduce the satellite snow cover data (the snow products, see section 2.3). In the case that multiple products were used for the evaluation of a model component and the construction of the posterior distributions of the associated parameters, a combined performance metric was formulated, based on the difference between 1 and the Euclidean distance (e.g., Fovet et al., 2015) between a vector of the model performances with respect to the individual products and a vector corresponding to perfect performance with respect to all objective functions (a vector of ones), thus treating the performance metrics equally important (equation (2)). Hereafter, rescaling was applied (equation (3)) to maintain values between 0 and 1.

$$E_{\text{obj,combined}} = 1 - \sqrt{\left(1 - E_{\text{obj},1}\right)^2 \ldots + \left(1 - E_{\text{obj},m}\right)^2} \tag{2}$$

$$L(\theta) = E_{\text{obj,combined,scaled}} = \frac{E_{\text{obj,combined}} - \min\left(E_{\text{obj,combined}}\right)}{\max\left(E_{\text{obj,combined}}\right) - \min\left(E_{\text{obj,combined}}\right)} \tag{3}$$

where $E_{\text{obj, combined}}$ is the combined objective function (or performance metric) using m remote sensing products for evaluation, $E_{\text{obj,m}}$ is the objective function value for product m, $E_{\text{obj,combined,scaled}}$ is the scaled objective, and $L(\theta)$ is an informal likelihood weight for parameter set θ. The posterior probability density function $L_2(\theta)$ was then determined as above (section 2.4) from equation (1).

This procedure was repeated for all possible combinations of remote sensing products from the four groups (Table 2), starting with a single product and ending with the combined use of all 10 products simultaneously. This resulted in a total of 1,023 different possible combinations of remote sensing products for the evaluation of the associated model components.

After weighting, the 25th and 75th quartiles of the posterior parameter distributions were retained as feasible parameter bounds. This remains a mere subjective choice, but in this way the higher values of the posterior likelihood are always retained and the new parameter ranges are always determined on a sufficient number of samples. The alternative of cutting off the distributions at a certain performance level, as for example done in the GLUE methodology (Beven & Freer, 2001), is equally subjective, with the additional risk of not having any feasible solutions to determine posteriors. The authors fully acknowledge that in absence of a clear posterior distribution the new bounds may incorrectly consider some solutions unfeasible (i.e., false negatives),

Table 3
Model Parameters and the Data Sources That Are Related to it for Determining Posterior Parameter Bounds

	Source	FLEX	FLEXtopo	HYPE	HYMOD	TUW
Soil Moisture	AMSR-E	Mmelt	Mmelt	Wcfc	Ts	Csf
		Tthresh	Tthresh	lp	Cfmax	Ddf
		Imax	Imax_p	mactrinf	CFR	Tr
		Sumax	Imax_h	Ttpd	CWH	Ts
		Beta	Imax_w	Ttpi	Sm	Meltt
		Kf	Sumax_p	Ttmp	Beta	FC
		Ks	Sumax_h	Cmlt	Alfa	BETA
		D	Sumax_w	Rrcs2	Rs	lprat
				macrate	Rf	K2
						cperc
	ASCAT	Mmelt	Mmelt	Wcfc	Ts	Csf
		Tthresh	Tthresh	lp	Cfmax	Ddf
		Imax	Imax_p	mactrinf	CFR	Tr
		Sumax	Imax_h	Ttpd	CWH	Ts
		Beta	Imax_w	Ttpi	Sm	Meltt
		Kf	Sumax_p	Ttmp	Beta	FC
		Ks	Sumax_h	Cmlt	Alfa	BETA
		D	Sumax_w	Rrcs2	Rs	lprat
				macrate	Rf	K2
						cperc
	NDII	Mmelt	Mmelt	Wcfc	Ts	Csf
		Tthresh	Tthresh	lp	Cfmax	Ddf
		Imax	Imax_p	mactrinf	CFR	Tr
		Sumax	Imax_h	Ttpd	CWH	Ts
		Beta	Imax_w	Ttpi	Sm	Meltt
		Kf	Sumax_p	Ttmp	Beta	FC
		Ks	Sumax_h	Cmlt	Alfa	BETA
		D	Sumax_w	Rrcs2	Rs	lprat
				macrate	Rf	K2
						cperc
	SMOS	Mmelt	Mmelt	Wcfc	Ts	Csf
		Tthresh	Tthresh	lp	Cfmax	Ddf
		Imax	Imax_p	mactrinf	CFR	Tr
		Sumax	Imax_h	Ttpd	CWH	Ts
		Beta	Imax_w	Ttpi	Sm	Meltt
		Kf	Sumax_p	Ttmp	Beta	FC
		Ks	Sumax_h	Cmlt	Alfa	BETA
		D	Sumax_w	Rrcs2	Rs	lprat
				macrate	Rf	K2
						cperc
Evaporation	Budyko	Imax	Imax_p	Lp	Sm	FC
		Sumax	Imax_h	Wcfc	Beta	BETA
		Beta	Imax_w	mactrinf		lprat
			Sumax_p			
			Sumax_h			
			Sumax_w			
	LSA-SAF	Imax	Imax_p	Lp	Sm	FC
		Sumax	Imax_h	Wcfc		BETA
		Beta	Imax_w	mactrinf		lprat
			Sumax_p			
			Sumax_h			
			Sumax_w			
	MOD16	Imax	Imax_p	Lp	Sm	FC
		Sumax	Imax_h	Wcfc		BETA
		Beta	Imax_w	mactrinf		lprat
			Sumax_p			
			Sumax_h			
			Sumax_w			
Total water storage	GRACE	Mmelt	Imax_p	Lp	Ts	Csf
		Tthresh	Imax_h	Wcfc	Cfmax	Ddf

Table 3 *(continued)*

	Source	FLEX	FLEXtopo	HYPE	HYMOD	TUW
		Imax	Imax_w	Mactrinf	Alfa	Tr
		Sumax	Sumax_p	Ttpd	Rs	Ts
		Beta	Sumax_h	Ttpi	Rf	Meltt
		Kf	Sumax_w	Ttmp		FC
		Ks	Ks	Cmlt		BETA
		D				lprat
						Cperc
						K2
Snow	MOD10	Mmelt	Mmelt	Ttpd	Ts	Csf
		Tthresh	Tthresh	Ttpi	Cfmax	Ddf
				Ttmp	CFR	Tr
				Cmlt	CWH	Ts
						Meltt
	MYD10	Mmelt	Mmelt	Ttpd	Ts	Csf
		Tthresh	Tthresh	Ttpi	Cfmax	Ddf
				Ttmp	CFR	Tr
				Cmlt	CWH	Ts
						Meltt

Note. See supporting information section S1 for a description of the models and the model parameters. AMSR-E = Advanced Microwave Scanning Radiometer - Earth Observing System;ASCAT = Advanced SCATterometer; NDII = Normalized Difference Infrared Index; LSA-SAF = Land Surface Analysis - Satellite Application Facility; GRACE = Gravity Recovery and Climate Experiment; SMOS = Soil Moisture and Ocean Salinity.

but therefore, the weighting factor n was set relatively high ($n = 10$). In this way, and in combination with the use of the 25th and 75th quartiles as bounds, considerable discriminative power can be obtained, zooming in on the solutions that are considered correct. All the possible combinations of products should therefore lead to a set of model results that are obtained without the use of observed discharged data. The method is thus used as if there is no discharge data available, and discharge data are not used to reject any model.

2.6. Parameter Selection—Benchmarking Streamflow Performances

All five models for all 27 study catchments were assessed for their performance range on observed streamflow data to provide a reference benchmark. Thus, all model realizations were evaluated against streamflow with a multiobjective strategy based on the Nash-Sutcliffe efficiency (E_{NS}) of flow and the Nash-Sutcliffe efficiency of the logarithm of the flow (E_{NSlog}). The 80,000 (100,000) samples obtained by Latin Hypercube sampling were used here as well. Model runs were now maintained as feasible when both E_{NS} and E_{NSlog} were higher than 0. This was preferred over calibration with, for example, automated optimization schemes, such as Shuffled Complex Evolution algorithm (Duan et al., 1992) or Dynamically Dimensioned Search algorithm (Tolson & Shoemaker, 2007), as sets of multiple *feasible* rather than one *optimal* parameter combination were sought. In this way, the benchmarking strategy on streamflow and constraining on remotely sensed data both generate empirical distributions of performances, which makes a fair comparison possible. Also here, the two objective functions were combined into a vector and the difference between 1 and the Euclidean distance to perfect performance for this vector is used (equation (4)):

$$E_{\text{obj,combined,streamflow}} = 1 - \sqrt{(1 - E_{NS})^2 + (1 - E_{NSlog})^2} \quad (4)$$

where $E_{\text{obj,combined, streamflow}}$ is the combined objective for streamflow.

2.7. The Added Value of Remote Sensing Data to Reproduce Streamflow

The added value of the individual remote sensing products was assessed by computing the Kolmogorov-Smirnoff test statistic for improvement when a specific data source is included (see section 2.3) to constrain the feasible parameter space of a given model, compared to not including this specific data source. High values of improvement correspond thus to high, positive values of the KS statistic. In addition, improvement was considered here relative to $E_{\text{obj, combined}}$ as defined in equation (5).

To do so, all 1,023 possible combinations of the 1 to 10 potential data sources from the four groups specified in section 2.3 were separated in combinations *with* and *without* a specific product, leading to 512 combinations *with* and 511 combinations *without* this product. Each combination has its own set of feasible solutions (as derived by the posterior parameter distributions) with the associated performance measures (such as E_{NS} or E_{NSlog}). The overall improvement with respect to a performance measure of including one product is then estimated by merging the distribution of performance measures for each combination *with* a specific product with the distributions of performance measures of all other combinations *with* this product into a combined empirical distribution of performances when using this product (Figure 2a). For example, if combinations A, B, and C each included the remote sensing product GRACE, with 100, 200, 250 feasible solutions, respectively, the final set contained 550 feasible solutions (100 + 200 + 250). Following the same approach for all combinations *without* this product, a second combined empirical distribution of performances is established.

The Kolmogorov-Smirnoff two-sample statistic (D^+) between the two empirical distributions can now be calculated and tested for significance. Thus, similar to the above, but in a more formal way, for each combination of satellite products c_j we have now selected a set of feasible parameters:

$$G_{c_j} \tag{5}$$

We define the union of all sets of feasible parameters selected, when using a given satellite product p to constrain:

$$G_p = \cup_{j=1,p\in c_j}^{1023} G_{c_j} \tag{6}$$

and the union of all sets of parameters selected when not using a given satellite product to constrain the parameter set:

$$B_p = \cup_{j=1,p\notin c_j}^{1023} G_{c_j} \tag{7}$$

We then apply $E_{obj,combined}$ to both sets of parameters, this results in two sets of scores S_{with} and $S_{without}$. For both sets we determine the frequency distribution. Next we test the following null and alternative hypothesis:

H0 S_{with} *comes from the same distribution as* $S_{without}$.
H1 S_{with} *is stochastically larger than* $S_{without}$.

2.8. The Added Value of Remote Sensing Data to Reproduce Hydrological Signatures

To assess the potential of using different remote sensing products to improve the representation of hydrological signatures compared to those obtained in an unconstrained situation, the feasible solutions of the parameter selection procedure as described in section 2.4 were evaluated for a set of 27 hydrological signatures (Table 4), as previously defined by, among others, Shamir et al. (2005), Yilmaz et al. (2008), Euser et al. (2013), Pechlivanidis and Arheimer (2015), and Kuentz et al. (2017). The Kolmogorov-Smirnoff two-sample D^+ statistic in the representation of hydrological signatures when including a specific remote sensing product was determined in comparison with a reference situation, which in this case corresponds to the distribution of signatures obtained from the unconstrained models (Figure 2b). We apply the performance metrics for a specific signature in the constrained case to get a set of values $S_{constrained}$ and in the unconstrained case to get a set of values $S_{unconstrained}$. For both sets we determine the frequency distribution. Next, we test the following null and alternative hypothesis:

H0 $S_{constrained}$ *and* $S_{unconstrained}$ *follow the same frequency distribution.*
H1 $S_{constrained}$ *and* $S_{unconstrained}$ *have different frequency distributions and* $S_{constrained}$ *scores are stochastically larger than* $S_{unconstrained}$ *scores.*

The ability of the models to reproduce the signatures was determined by the Nash-Sutcliffe efficiency (E_{NS}) between observed and modeled signatures in case of (time) series, only for single-valued signatures the

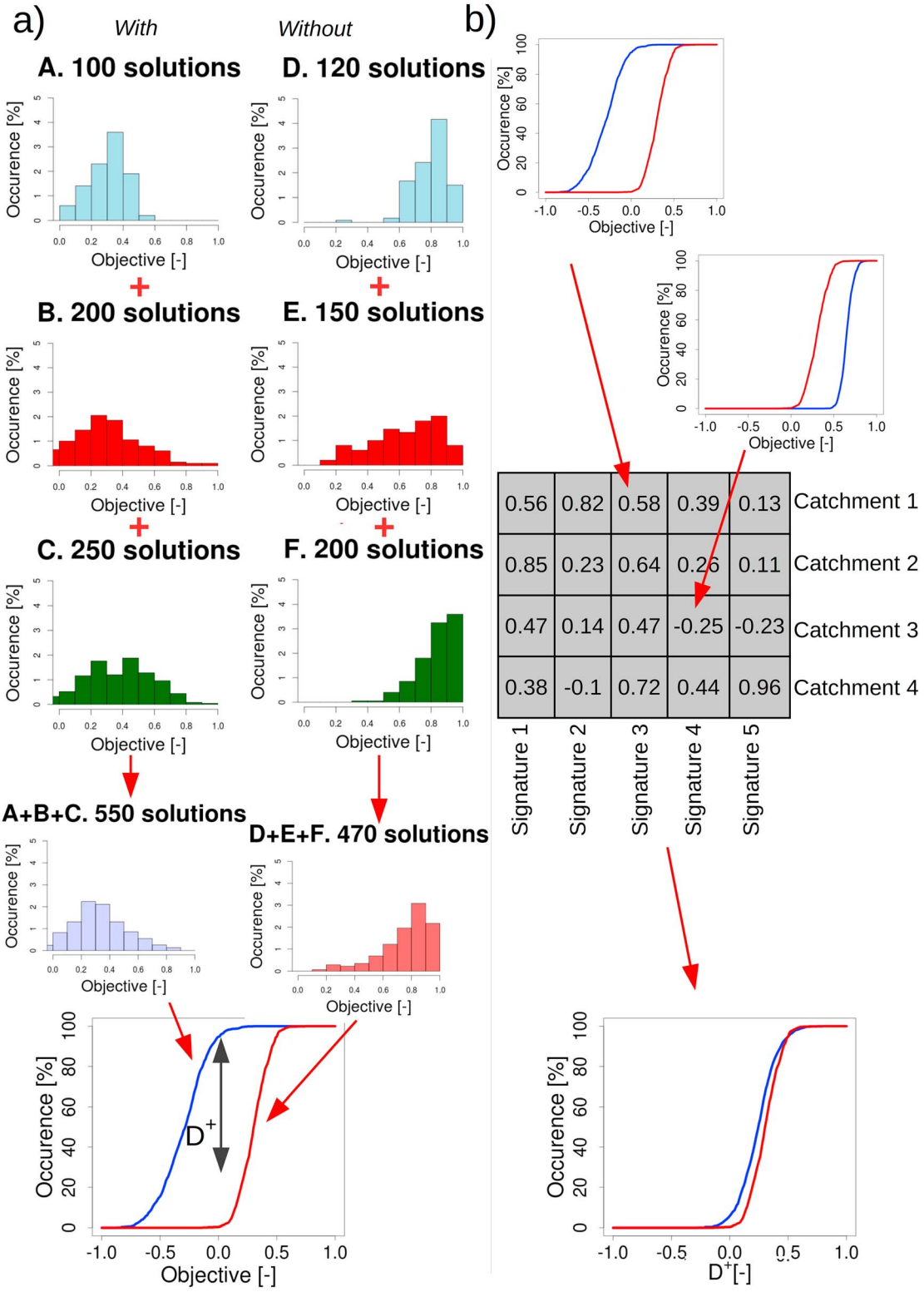

Figure 2. Schematized representation of (a) the procedure to calculate the Kolmogorov-Smirnoff test statistic D^+ for each product by adding up all feasible solutions in Step 1 *with* a certain product, as shown for examples A (100 feasible solutions), B (200), and C (250), which leads to a marginal distribution in Step 2 of 550 solutions (100 + 200 + 250) and *without* a certain product, as shown for the cases D, E, F (120 + 150 + 200 = 470 solutions), which can be compared in Step 3 to calculate the test statistic D^+ between the empirical distribution curves. In Figure 2b the signature analysis is displayed, with in Step 1 the frequency distributions for the performance metrics for a specific signature in the constrained case (red) and unconstrained case (blue), Step 2 the KS statistics derived from these distributions for each signature and each catchment, and Step 3 the resulting cumulative occurrences for these statistics.

Table 4
Hydrological Signatures Applied in the Signature Analysis

Signature	Description	Reference
S_{QMA}	Mean annual runoff	
S_{AC}	One day autocorrelation coefficient	Montanari and Toth (2007)
$S_{AC,summer}$	One day autocorrelation the summer period	Euser et al. (2013)
$S_{AC,winter}$	One day autocorrelation the winter period	Euser et al. (2013)
S_{RLD}	Rising limb density	Shamir et al. (2005)
S_{DLD}	Declining limb density	Shamir et al. (2005)
S_{Q5}	Flow exceeded in 5% of the time	Jothityangkoon et al. (2001)
S_{Q50}	Flow exceeded in 50% of the time	Jothityangkoon et al. (2001)
S_{Q95}	Flow exceeded in 95% of the time	Jothityangkoon et al. (2001)
$S_{Q5,summer}$	Flow exceeded in 5% of the summer time	Yilmaz et al. (2008)
$S_{Q50,summer}$	Flow exceeded in 50% of the summer time	Yilmaz et al. (2008)
$S_{Q95,summer}$	Flow exceeded in 95% of the summer time	Yilmaz et al. (2008)
$S_{Q5,winter}$	Flow exceeded in 5% of the winter time	Yilmaz et al. (2008)
$S_{Q50,winter}$	Flow exceeded in 50% of the winter time	Yilmaz et al. (2008)
$S_{Q95,winter}$	Flow exceeded in 95% of the winter time	Yilmaz et al. (2008)
S_{Peaks}	Peak distribution	Euser et al. (2013)
$S_{Peaks,summer}$	Peak distribution summer period	Euser et al. (2013)
$S_{Peaks,winter}$	Peak distribution winter period	Euser et al. (2013)
$S_{Qpeak,10}$	Flow exceeded in 10% of the peaks	
$S_{Qpeak,50}$	Flow exceeded in 50% of the peaks	
$S_{Qsummer,peak,10}$	Flow exceeded in 10% of the summer peaks	
$S_{Qsummer,peak,50}$	Flow exceeded in 10% of the summer peaks	
$S_{Qwinter,peak,10}$	Flow exceeded in 10% of the winter peaks	
$S_{Qwinter,peak,50}$	Flow exceeded in 50% of the winter peaks	
S_{SFDC}	Slope flow duration curve	Yadav et al. (2007)
S_{LFR}	Low flow ratio (Q_{90}/Q_{50})	
S_{FDC}	Flow duration curve	Westerberg et al. (2011)

relative error was used (E_{RE}). The Kolmorov-Smirnoff statistic D^+ with respect to the performance metrics of the signatures (the Nash-Sutcliffe efficiency and relative error) can be calculated for each signature and each catchment.

The Kolmogorov-Smirnoff statistic D^+ is calculated for all 729 cases (27 catchments with 27 signatures each) for each combination of data sources (i.e., a certain set of constraints), with the improvement relative to the unconstrained reference model. The relative occurrences of certain, significant, KS statistics can be inspected by means of cumulative frequency plots. See also Figure 2b for a stepwise clarification of this approach. As the different combinations of products lead to a varying degree of signature reproduction, different cumulative frequency curves will emerge. Therefore, improvements can be identified by shifts in the cumulative frequency plots toward higher, positive values of the KS statistic D^+.

3. Results and Discussion

3.1. Linking Parameters and Data Sources

In a first step, all parameters had to be related to relevant data sources. This was done based on the sensitivity for each parameter to a certain product (Figure 3), in terms of the maximum vertical distance, averaged over the catchments, between the prior cumulative distribution and the posterior cumulative distribution (Kolmogorov-Smirnoff statistic; i.e., the higher the distance, the more sensitive a parameter is to the information provided by a given product). For FLEX (Figure 3a), it can be noted that the two snow parameters (*Mmelt* and *Tthresh*) react to the two snow products, as expected. More interestingly, also the soil moisture products and GRACE influence the snow parameters, which can be explained by the role of snowmelt filling up the unsaturated storage and thus generating runoff. A similar argumentation holds for the parameter of maximum interception capacity *Imax*, which is, in addition to solely the evaporation products, also affected by the soil moisture products. The soil moisture parameters S_{umax} and *Beta* exhibit a high sensitivity to the

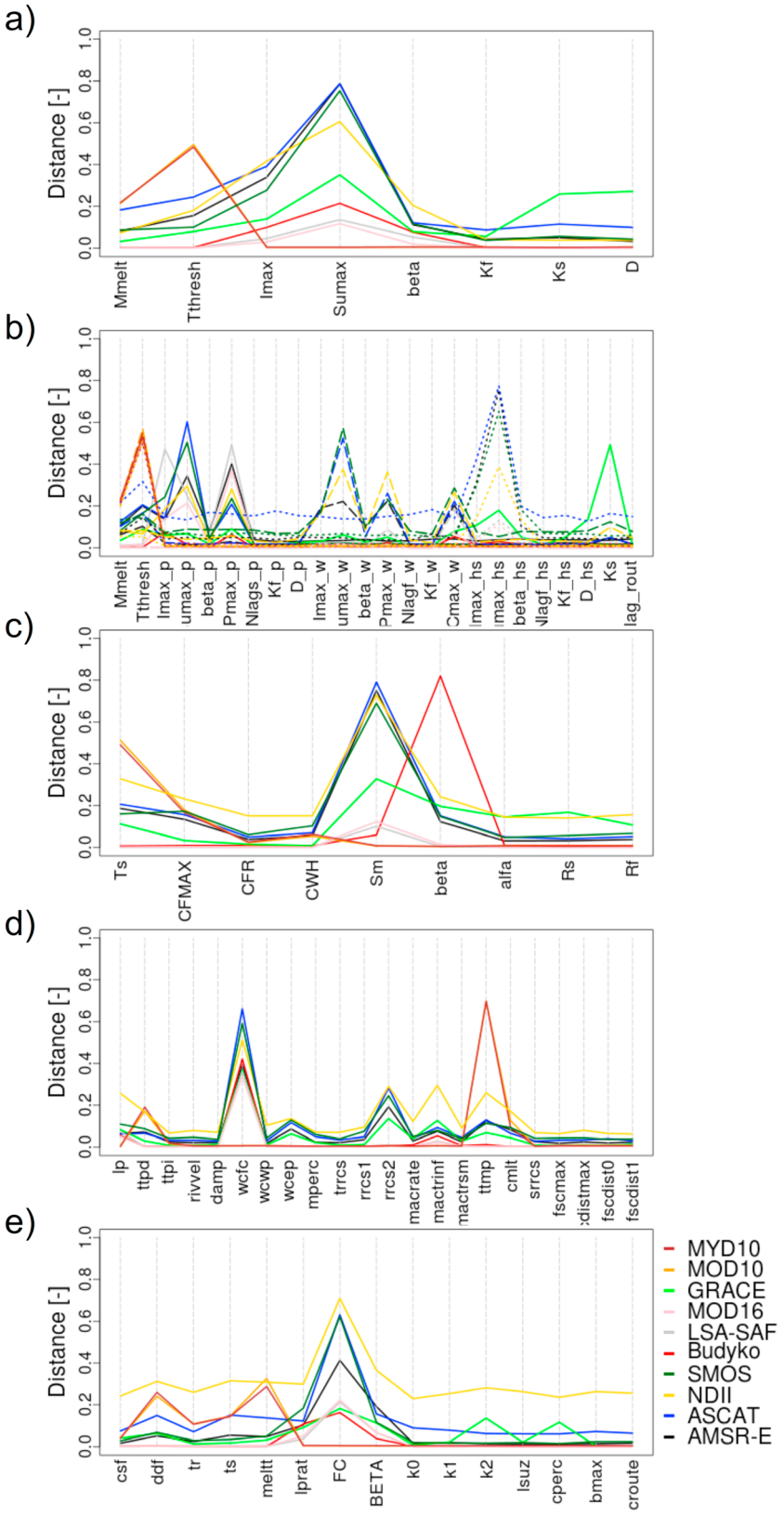

Figure 3. Sensitivity in average vertical distance between the empirical distribution curves of posterior and prior (uniform) parameter distributions for (a) FLEX, (b) FLEXtopo, (c) HYMOD, (d) HYPE, and (e) TUW. Different colors indicate different products, with for FLEXtopo a distinction per landscape class with plateau, wetland (dashed), and hillslope (dotted).

group of soil moisture products and also to the evaporation products and GRACE. The parameters *Kf*, *Ks*, and *D* are to some extent linked to GRACE and also to the soil moisture products. Thus, each parameter of FLEX can be related to certain products and can hence be constrained.

The sensitivity plot for FLEXtopo (Figure 3b) shows similar behavior, as similar model parameters are used as in FLEX, but now only applied in different landscape classes. Even though no complete overlap could be found, corresponding parameters in FLEXtopo were constrained with the same products as for FLEX to maintain consistency. However, as the number of parameters is higher, while the number of sampled random parameter sets is just slightly higher, the analysis of this model is based on a lower sampling density. Thus, to avoid the situation that no solutions remain due to too many parameter constraints, parameters that did not show a consistent sensitivity for a specific group of products were mainly left unconstrained, such as *Pmax*, *Cmax*, *Kf*, *Beta*, and *D* for the three landscape classes (Figure 3b).

The relations between products and parameters for HYMOD (Figure 3c) show a rather consistent pattern compared to FLEX and FLEXtopo. Also in this case, the snow parameters (*Ts*, *CFMAX*, *CFR*, *CWH*) are not only sensitive to the snow products but also the soil moisture products. For *Ts* and *CFMAX* also GRACE has an influence, which happened as well for the snow parameters of FLEX. *CFR* and *CWH* do not show this, indicating that this refreezing factor (*CFR*) and water holding capacity of snow (*CWH*) have, apparently, a minor influence in the storage anomalies. The maximum soil moisture *Sm* can be constrained with the soil moisture products, GRACE, and the evaporation products. For the second soil moisture parameter *beta,* the soil moisture products matter, but especially a high sensitivity to the Budyko framework can be observed. The last parameters of *Rs*, *Rf*, and *alfa* (relating to the reservoir coefficients and runoff generation) can be linked to the soil moisture products as well as GRACE.

It can be noted for HYPE (Figure 3d) that the general sensitivity is lower compared to the other models. Nevertheless, the snow parameters (*ttpd*, *ttpi*, *ttmp*, and *cmlt*) can be constrained in a similar fashion as for FLEX and HYMOD, thus with the snow products, soil moisture products, and GRACE. In addition, *rrcs2* and *macrate*, both related to groundwater dynamics, show considerable sensitivity to GRACE and the soil moisture products. The parameters controlling soil moisture and, thus transpiration *lp*, *wcfc*, and *mactrinf*, relate to GRACE, the soil moisture products, and the evaporation products. The other parameters are left unconstrained as the sensitivities here are generally low or do not show a clear sensitivity to one of the groups of products.

The snow parameters *csf*, *ddf*, *tr*, *ts*, and *meltt* of the TUW model (Figure 3e) can also be constrained with the snow products, GRACE, and the soil moisture products. The soil moisture parameters of *lprat*, *FC*, and *BETA* show a high sensitivity to the soil moisture products, and also again GRACE and the evaporation products. The groundwater parameters *k2* and *cperc* relate to GRACE and the soil moisture products, and the remainder of the parameters is left unconstrained as no clear preference for certain groups of products can be identified.

3.2. Benchmarking Streamflow Performances Versus Constrained Performances

In a next step, the range of performances of the five models when constrained with all different combinations of data sources (section 2.4) is compared to the performance range of the benchmark solutions, that is, the set of solutions with E_{NS} and $E_{NSlog} > 0$ (section 2.5). The models provide on average similar model performances, but different patterns emerge for the different catchments and models when constrained with the different data sources as summarized in Figure 4. Here the empirical cumulative distribution curves for Nash-Sutcliffe efficiencies of all parameter sets for all catchments retained as feasible are shown for the benchmarking situation, as well as for constraining the models with all 10 products and the unconstrained situation.

It can be noted that for FLEX (Figure 4a) many of the poorly performing solutions with very low Nash-Sutcliffe efficiencies of the unconstrained case are identified and discarded when constraining the model on 10 remote sensing products. In fact, the distribution of Nash-Sutcliffe efficiencies of the respective solutions retained as feasible remains similar for both, the benchmark and constraining on remotely sensed products, and that the higher values are maintained when constrained on the 10 products. Compared to the unconstrained situation, the remote sensed products provide enough information to shift the curve toward the benchmarking situation. In contrast, for FLEXtopo (Figure 4b) and HYMOD (Figure 4c), the curve after constraining with the products reaches lower maximum values compared to the benchmarking situation

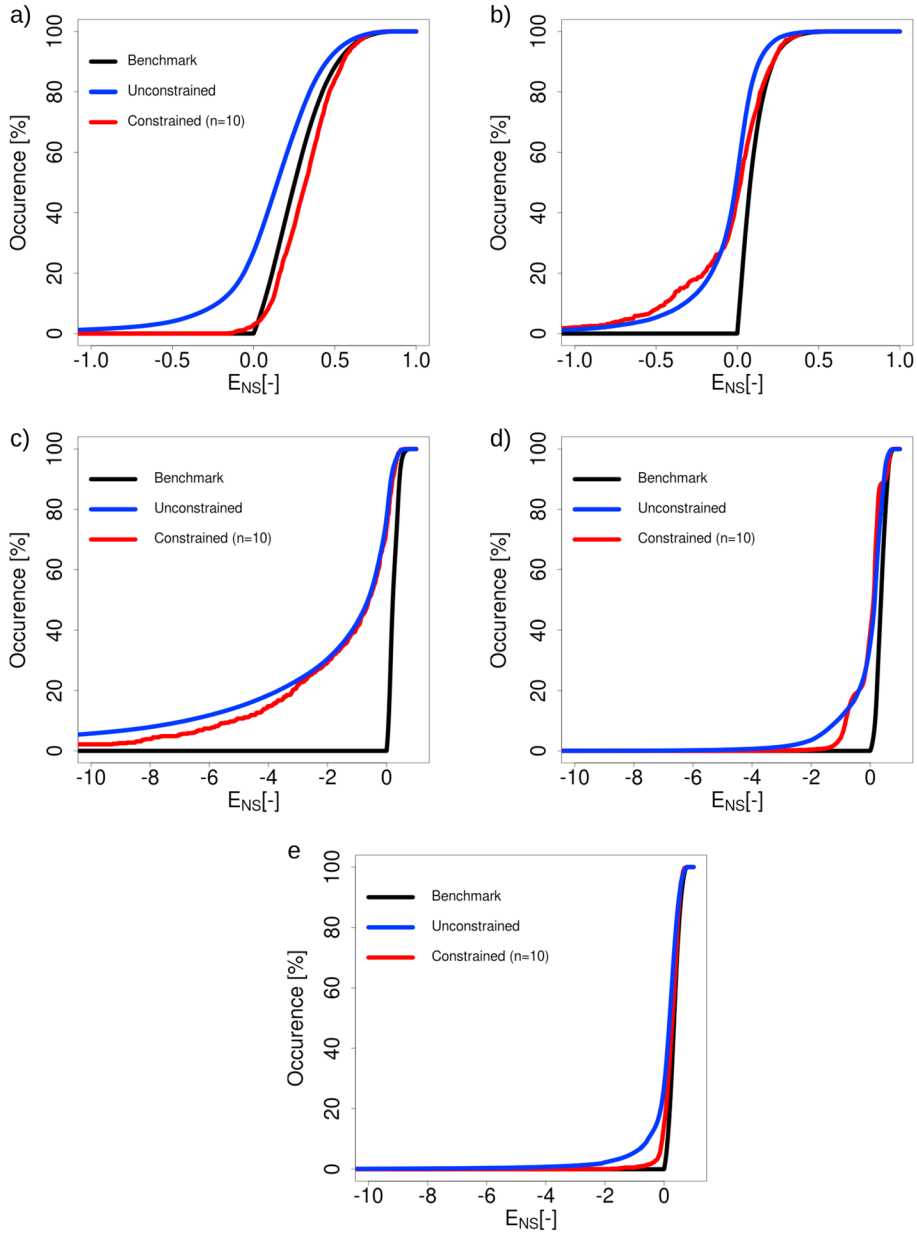

Figure 4. Empirical cumulative distribution curves of Nash-Sutcliffe efficiencies for (a) FLEX, (b) FLEXtopo, (c) HYMOD, (d) HYPE, and (e) TUW for all catchments in the benchmarking situation (black), constraining on remotely sensed data (red) and the unconstrained situation (blue).

(i.e., top of the curve remains left of the curve obtained by the benchmarking situation). Hence, several solutions that lead to high Nash-Sutcliffe values are discarded when constrained on the remotely sensed products, which indicates that the constraints are too restrictive and consider several solutions incorrectly as unfeasible. On the other hand, it may also point toward parameterizations that lead to high objective function values that are not hydrologically consistent or even deceptive (e.g., Andréassian et al., 2012; Kirchner, 2006). Nevertheless, more importantly, it can be noted in Figure 4c that the constraints help in filtering out the worst solutions for HYMOD, as can be seen when comparing the unconstrained curve with the curve obtained by the constrained situation. However, many poor solutions are maintained, with even Nash-Sutcliffe values less than zero. These solutions would definitely be discarded in a more traditional calibration, as this means that the model performs worse than the long-term mean of the observations. In absence of knowledge about these long-term statistics, being an exercise in predicting ungauged basins,

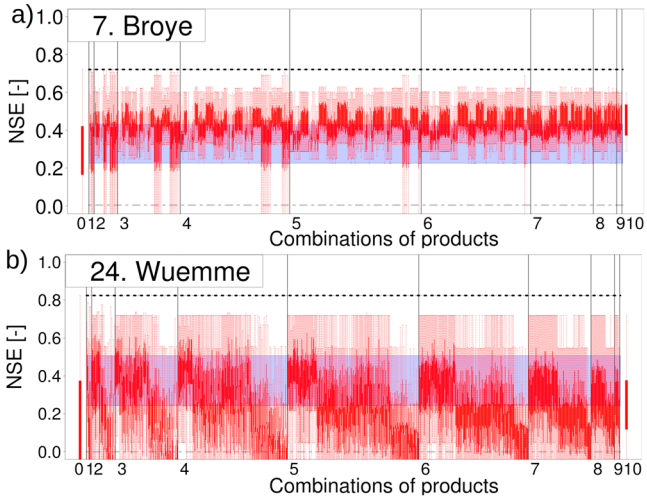

Figure 5. Comparison of Nash-Sutcliffe performances for the benchmark situation (25th and 75th quartile correspond to the blue band; dashed gray lines represent the 5th and 95th quantiles) and the constrained model applications (red boxplot for each combination of constraints) for (a) FLEX and the Broye catchment, and (b) TUW and the Wuemme catchment. The wider red box plots on the left (n = 0) and right (n = 10) represent the unconstrained and fully constrained situation respectively. The maximum Nash-Sutcliffe performance in the benchmark situation is indicated with the black, dashed line.

the solutions are however maintained here as a (unfortunate) result. The constraints are therefore not restrictive enough to zoom in on the high performances, but the constraints at least help in substantially narrowing the parameter search space by removing the very worst solutions. For HYPE and TUW, it can be seen in Figure 4 that the higher values are still maintained, but that, at the same time, relatively many solutions are maintained as feasible when constrained on the 10 products. This is also reflected by the close resemblance between the unconstrained and constrained curves. It can therefore be argued here that the constraints do not have enough discriminative power to closely zoom in on the high objective function values, but at least the solution space is reduced and does still contain the high Nash-Sutcliffe efficiencies. This lack of constraining power can also be related to the relatively large number of parameters for these two models of 22 (HYPE) and 15 (TUW).

The models show, however, a strong variable pattern in model objective functions over the full range of combinations used for constraining, as shown for selected catchments and models in Figure 5 and the supporting information section S2. Here the performance ranges of the benchmarking solutions for streamflow are shown for the 25th and 75th quartile (blue) and 5th and 95th quartile (i.e., a wide boxplot). Similarly, boxplots for each combination of constraints are shown in red, ranging from unconstrained (left) to using all 10 products for the constraints (right). In general, the model performances increase with an increasing number of data sources used for constraining, as expected, but in some occasions no feasible solutions remain (i.e., zero solutions are left after constraining). It can however still be noted in Figure 5 that constrained models still not reach the highest performances in the benchmarking situation.

In general, some model applications constrained exclusively by remote sensing data exhibit a similar range of model performances as the model performance ranges for the benchmark, but often the constrained models are still outperformed by the benchmarking situation. In several cases an equivalent level of model performance was achieved after adding four data sources in the parameter selection procedure. As an example, in Figure 6 the maximum distance between empirical cumulative distribution curves of the benchmark and constrained TUW model (the Kolmogorov-Smirnoff statistic) is displayed for all combinations of products and catchments, starting on top with a single product used for constraining (n = 1) toward using all 10 products (n = 10) at the bottom of the figure, for each catchment (x axes). In this figure it can be noted that groups of blue boxes and red boxes exist, pointing at groups with either a positive or negative test statistic D^+ between the benchmark and constrained empirical distributions. Similarly, in Figure 5, zigzag patterns of improvements and deteriorations can be noticed after constraining with a larger set of products for some catchments (moving toward the right on the x axes). This is an indication that families of combinations that either include or exclude certain (combinations of) products can lead to major improvements or strong decreases in performances. It can be clearly observed that the values vary and are grouped (Figure 6), pointing at specific combinations of products that constrain in such a way that the new performance distribution comes close to the benchmark distribution or even improves. However, it can also be noted in Figure 6 that many cases exist where no improvements are observed.

The TUW model (supporting information Figures S24–S26 and Figure 6a) shows also a clear pattern of strong and weak combinations of remote sensing products, but the variability between the (families of) combinations is generally not very high. The relative importance of each parameter, and thereby each data source connected to it, is lower as TUW has a relatively elevated number of parameters (i.e., 15). Thus, leaving a single parameter of all the TUW parameters unconstrained has less consequences compared to constraining a single parameter from the eight parameters of FLEX. Nevertheless, the same products (GRACE, AMSR-E, and ASCAT) strongly improve the parameterizations of TUW, similar to FLEX, also for the same catchments (such as catchments 8, 20, and 24).

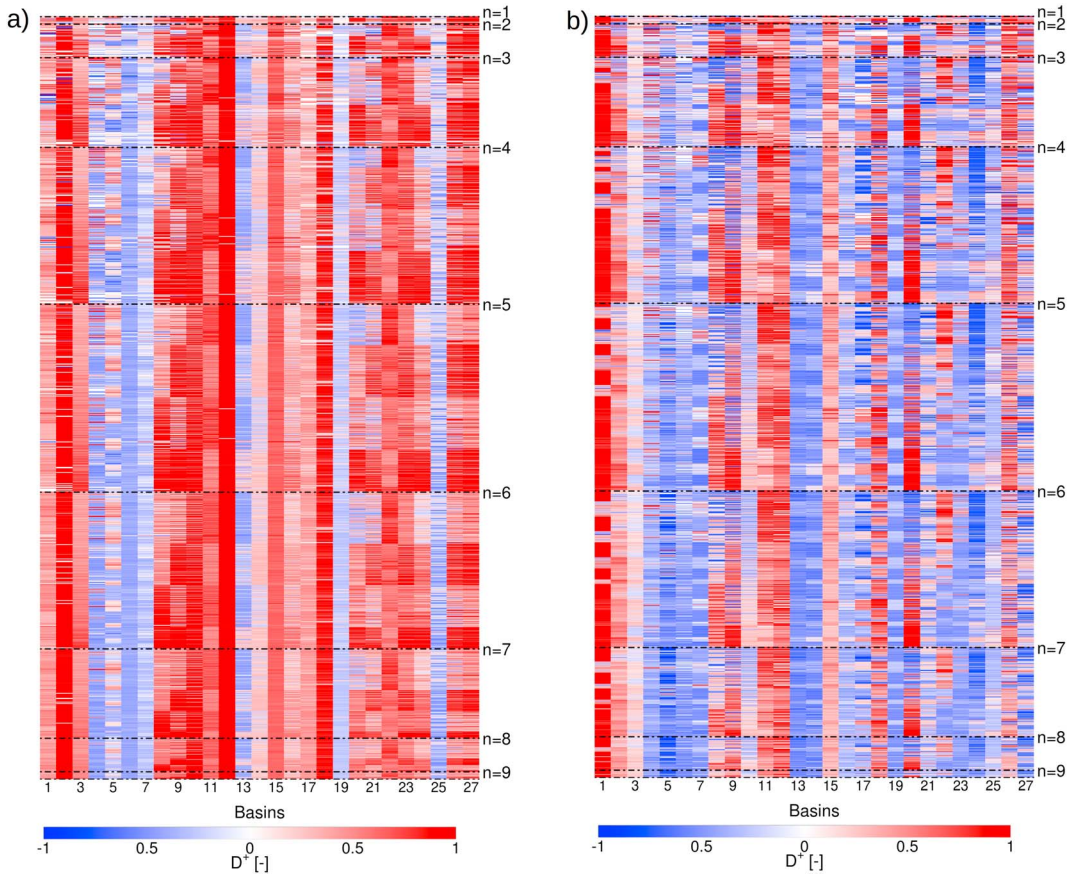

Figure 6. Distance between the benchmark and constrained empirical cumulative distributions of the Euclidean distance between Nash-Sutcliffe of the flows and log of the flows (i.e., Kolmogorov-Smirnoff statistic), for each possible combination from $n = 1$ to $n = 9$ included products, for (a) the TUW model and (b) the FLEX model. Blue values indicate a constrained distribution with higher performances than in the benchmark situation; red indicates a benchmark distribution with higher performances.

For FLEX (supporting information Figures S9–S11 and Figure 6b), only in 4 of the 27 catchments (catchments 1, 11, 12, and 18) the constraints lead to performance distributions that are substantially lower than the benchmark results, while the other catchments always approach the benchmark model results more closely or even show higher performances (catchments 5, 13, 14, and 24). This variability between strong and weak parameterizations can also be clearly seen in Figure 6b, which is now bluer, indicating more strong parameterizations and improvements. In comparison with the other models, FLEX shows a stronger reduction of the parameter search space, when constrained with the remotely sensed products (Figures 4 and 6b). This is probably caused by the relatively low number of parameters, in combination with a rather simple and generally applicable model structure. Inspection of the combinations reveals that GRACE data are an important contributor to model improvements particularly for a large number of catchment basins, except for the Gadera (catchment 1), but especially for catchments 7, 14, 17, 19, and 27. Similarly, for the Treene, Modau, and Wuemme (catchments 8, 20, and 24) the AMSR-E product is the common factor in the more successful combinations. The Treene catchment and, to a lesser extent, also the Wuemme are peaty lowland catchments, with very moist soils and shallow groundwater tables, which match well with the information derived from AMSR-E. This seems, however, in contrast to statements from the AMSR-E developers that pixels with a large proportion of open water introduce errors (Owe et al., 2008) or other researchers that suggest that peatland creates errors in soil moisture products (Bartalis et al., 2007). On the other hand, Owe et al. (2008) mention steep mountainous areas as source for error, which these catchments are certainly not. The snow products are included in the more successful combinations for Vils, Grossarler, and Große Mühl (catchments 4–6), which are also the more snow-dominated catchments. The evaporation products of MOD16 and LSA-

SAF do not show a clear pattern when included or excluded in the parameter selection procedure, pointing at a relatively minor role in determining the performances with regard to streamflow, reflecting results of Oudin et al. (2004). In addition, the number of parameters concerning evaporation is generally lower, also reducing the importance of these products for constraining.

The results of the FLEXtopo model (supporting information Figure S12–S15) show in general similar patterns as for FLEX. However, it can be noted that more combinations of constraints also lead to no remaining feasible solutions at all. The problem here is mostly linked to the larger number of free parameters (24) and the resulting undersampled parameter space compared to FLEX. In addition, a similar reasoning can be made for HYMOD (supplementary material Figures S16–S19), but here this is merely caused by the relatively wide prior parameter ranges used for HYMOD, leading to only a relatively small number of solutions with high performances that the constraints cannot easily filter out. These ranges were initially set wide on purpose to assess the power of the constraints, but it largely remains a challenge to obtain similar performances as for the benchmark model, when set too wide. Thus, similar performances as for the benchmark situation are eventually only achieved for the Tanaro, Fyllean, and Deveron (catchments 2, 18, and 25). The variability between the different combinations of constraints is also large in the case of HYMOD, pointing to an extremely high added value of a certain (family of) constraints, which are combinations with the Budyko framework.

Similarly, HYPE (supporting information Figures S20–S23) has some more difficulties in order to obtain similar performances as for the benchmark situation. Even though the benchmarking performance ranges are reasonable, the absolute number of feasible solutions is relatively low. Therefore, the constraints from the products need considerably more restrictive power to filter all solutions and to converge to the same performance level as the benchmark situation. This relates also to the relatively large number of free parameters and thus a larger a priori search space. In other words, too many poor solutions are maintained when the model is constrained on the remote sensing data sources.

3.3. The Added Value of Specific Remote Sensing Data to Reproduce Streamflow

Figure 7 compares the overall added value of each individual remote sensing product to generate meaningful posterior distributions and thus to provide efficient and effective parameter constraints. This was done as described in section 2.6 by assessing the Kolmogorov-Smirnoff test statistic D^+ between the representation of streamflow when including a specific remote sensing data source for constraining a model compared to not including it. Figure 8 therefore reports the *p*-values obtained by this test. Additionally, Figures S6–S8 also show the performances in different ways. In Figure S6 and S7 the performances are ordered according to different catchment characteristics, and Figure S7 shows here the deseasonalized performances. In Figure S8 the performances of the remotely sensed products are plotted against the performances for streamflow.

For example, in Figure 7a it can be noted that strong improvements are obtained for most data sources for FLEX. Besides, most of the results are significant, as shown by the low *p*-values in Figure 8a. Only a few cases exist where one of the data sources provides a negative KS statistic, which suggests that constraining on any single remote sensing data source has already considerable constraining power in a wide range of cases. This is for example true for GRACE, in particular when applied with FLEX, where high KS statistics can be observed (see also section 3.2). This is similarly illustrated for the hydrograph of the Glane (catchment 13) in Figure 9, where especially in combinations with nine products GRACE has the potential to move the lower bound of the uncertainty interval of the modeled hydrograph toward the observations, in particular for the low flows. Further, it can be noted that during the summer period the final set of constraints provides a much narrower uncertainty bound than the benchmark results (Figure 9c). This may seem counterintuitive at first, but constraining on nine products is much more restrictive compared to two objective functions in the benchmarking strategy. In addition, in the benchmarking situation relatively many solutions are kept as feasible, as all solutions with E_{NS} and E_{NSlog} higher than zero are maintained. Hence, the envelopes here represent generally also the number of solutions that are kept as feasible. Nevertheless, several previous studies similarly suggested that GRACE has a high potential to improve hydrological simulations (Mulder et al., 2015; Rakovec, Kumar, Attinger, et al., 2016), but this was thought to be true mostly for larger catchments than those under consideration here. In particular, the Broye and Dalsaelve (catchments 7 and 17) show a strong improvement when constraining FLEX with GRACE. For the Broye this is likely to be linked to the adjacent Lake Geneva,

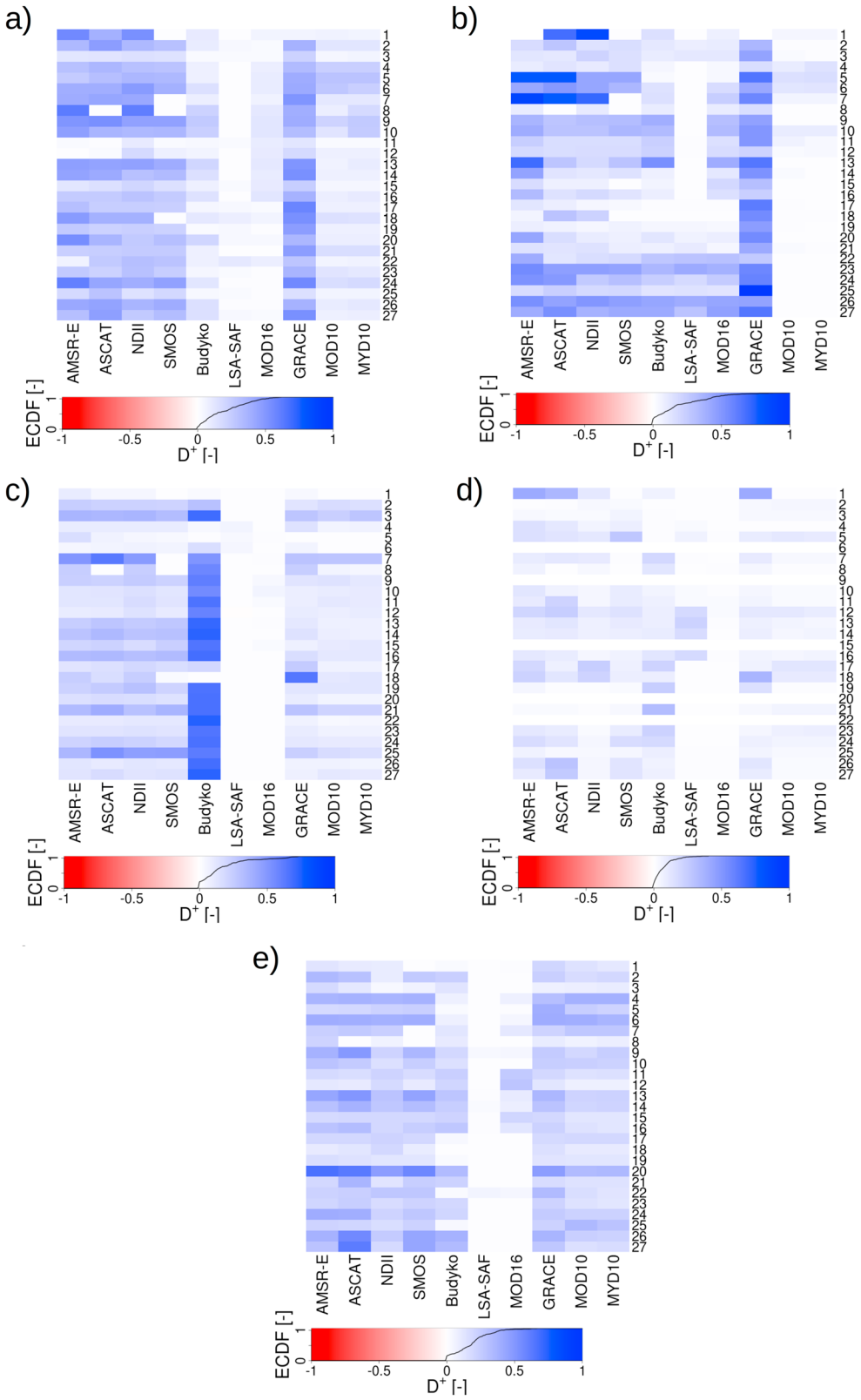

Figure 7. Kolmogorov-Smirnoff statistic (D^+) for including a specific product in the set of products used for constraining compared to not including this product, with regard to the Euclidean distance between Nash-Sutcliffe of flows and logarithm of the flows, shown for (a) FLEX, (b) FLEXtopo, (c) HYMOD, (d) HYPE, and (e) TUW. High values of D^+ plot in increasingly dark shades of blue, while shades of red indicate negative D^+. The curves represent the empirical cumulative distribution curves of all KS statistics for all catchments.

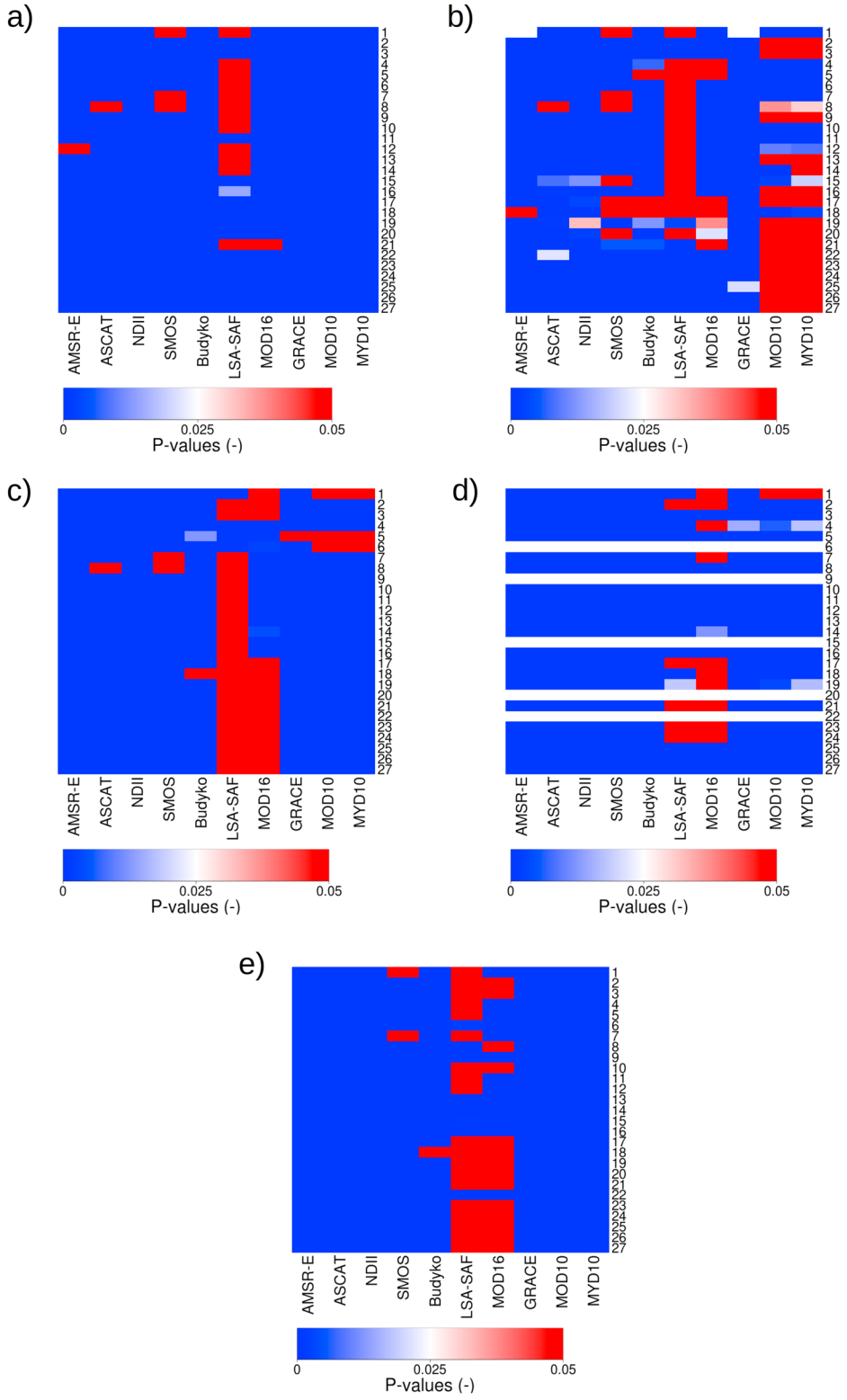

Figure 8. *P*-values for including a specific product in the set of products used for constraining compared to not including this product, with regard to 1 minus the Euclidean distance between Nash-Sutcliffe of flows and logarithm of the flows, shown for (a) FLEX, (b) FLEXtopo, (c) HYMOD, (d) HYPE, and (e) TUW. High *p*-values plot in increasingly dark shades of blue, toward red, values higher than 0.05 plot as red too. AMSR-E = Advanced Microwave Scanning Radiometer - Earth Observing System; ASCAT = Advanced SCATterometer; NDII = Normalized Difference Infrared Index; LSA-SAF = Land Surface Analysis - Satellite Application Facility; GRACE = Gravity Recovery and Climate Experiment; SMOS = Soil Moisture and Ocean Salinity.

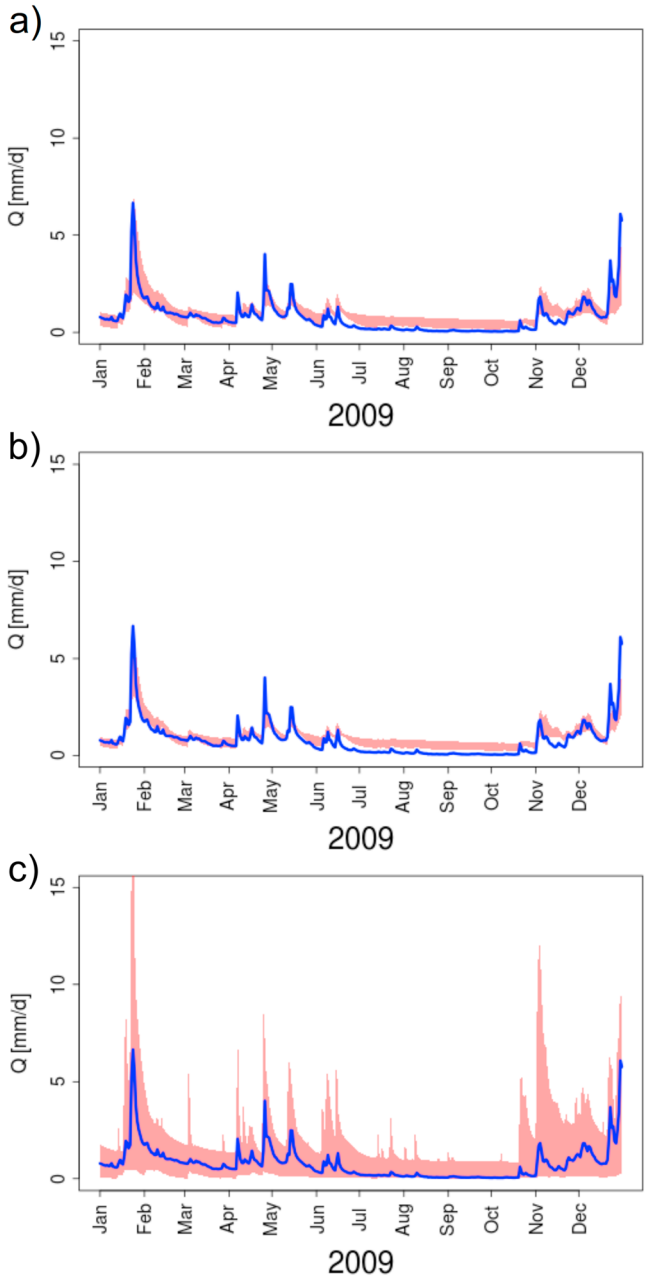

Figure 9. Feasible flow ranges for FLEX for a selected time period of catchment 13 obtained with combinations of nine remote sensing products (a) *with* GRACE and (b) *without* GRACE (c) the benchmarking situation. Colored envelopes in Figures 9a and 9b represent the number of products used in deriving the posterior parameter distributions and flow ranges, observed discharge is shown in blue and, in Figure 9c, discharge in the benchmark situation in red. GRACE = Gravity Recovery and Climate Experiment.

which may influence the groundwater tables in the surrounding catchments, leading to similar water storage anomalies of all catchments within the GRACE cells. The larger catchments, such as the Leyre (10; 1587 km^2) or Hunte (23; 1409 km^2) still show relatively high values for the KS statistic D^+, which is related to a signal of water storage anomalies much closer to the GRACE signal. Yet, the KS statistics D^+ are high for most other catchments as well, which also includes catchments with areas of less than 100 km^2 (e.g., catchment 20). In addition to GRACE, the soil moisture products of AMSR-E and ASCAT show the strongest signal of improvement, whereas SMOS has a slightly lower added value. This is in agreement with the findings of Wanders et al. (2014), who also applied AMSR-E, ASCAT, and SMOS in a hydrological model and found that soil moisture improved the strongest for AMSR-E and ASCAT. Nevertheless, they also found that AMSR-E and ASCAT work best in areas with a pronounced relief, whereas the highest KS values are, in our study, obtained for both catchments with low- and high-elevation differences. The relatively white colors for Budyko, MOD16 and LSA-SAF in Figure 7, corresponding to a relatively low difference between the two empirical distributions, indicate that these data sources do not add a lot of constraining power and also do not have adverse effects when included. These data sources are, apparently, not very important with respect to the streamflow objectives considered here. In addition, it can also be noted in Figure 8 that the results for LSA-SAF are often not significant, showing that including the product does not change the posterior distribution of performances. However, in warmer and more arid climates outside Europe, these products may have significantly more value. The snow products show, as expected, strong improvements for Vils, Grossarler, and Große Mühl (catchments 4–6), which are more snow-dominated catchments.

FLEXtopo (Figure 7b) shows a similar pattern as for FLEX, as it can also be observed that LSA-SAF evaporation has no significant influence (Figure 8b). This is in line with Figure S6 in the supporting information, where it can be noted that FLEXtopo has difficulties to achieve high correlations between the LSA-SAF evaporation and modeled evaporation, therefore also leading to rather poor posterior parameter ranges. It is also interesting to note that, in Figure 8b, the snow products only produce significant results in the more snow-dominated catchments, whereas for FLEX, even though with low values of the KS statistics, the results are still significant.

Unlike the results for FLEX and FLEXtopo, the Budyko framework has a big, significant influence on the results for HYMOD (Figures 7c and 8c). Here the high values for the KS statistics D^+ show the importance of the Budyko framework for HYMOD, whereas FLEX and FLEXtopo (and also HYPE and TUW) show an almost white column for Budyko. This indicates that the model has difficulties in reproducing the long-term flux partitioning into streamflow and evaporative fluxes, which can also be seen from Figure S6 in the supporting information. Nevertheless, the Budyko framework was only connected to two parameters (Sm and Beta) of HYMOD, and therefore, leaving these parameters unconstrained leads to many poor solutions. Based on these results, it can be argued that these parameters must be constrained in all cases. However, the Budyko framework helped here, similar as in the studies of Li et al. (2014) and Gentine et al. (2012), to identify feasible sets of parameters. Another clear distinction with the other models can be found in the KS values obtained for the two snow products (MOD10 and MYD10). For all other models at least moderate values are observed, in particular for catchments 4–7, but for HYMOD the differences between the two empirical distributions remain very low or are not even

significant (catchment 5 and 6, Figure 8c). This can also be noted from the hydrographs in Figure 10, where for FLEX some snow peaks are improved when constraining the snow parameters, for HYMOD the small snow peaks in January and February disappear, and the large snow peak starting in March remains too high.

Similar to FLEX and FLEXtopo, HYPE (Figure 7d) benefits from including ASCAT or AMSR-E, even though the absolute level of improvements remains quite low. NDII shows even lower values for all catchments. It is also interesting to note that the snow products have a similar low range in KS statistics for the Vils and Grossarler catchment (catchments 4 and 5) as for HYMOD, whereas FLEX, FLEXtopo, and TUW have more distinct difference between the benchmark and constrained model for applying the snow products in these catchments. These two catchments have the highest number of possible snow days (29.8% and 28.7%), and one would expect high-added value for the snow products here for all models, also based on previous work (Parajka & Blöschl, 2008). Figure S6 also shows that there are no distinct differences between the modeled and observed snow signals between the models and, thus, the low KS values, with regard to streamflow, very likely point toward other model structural deficiencies in HYMOD and HYPE. In other words, the snowmelt may still be better represented when the snow parameters are constrained with the snow products, but how snowmelt water is then routed through the rest of the system may not be adequately represented.

For TUW, the observed patterns in Figure 7e are again similar to the patterns for FLEX and FLEXtopo, but with less distinct differences between the different data sources. ASCAT also shows a high-added value for most catchments, and AMSR-E to a lesser degree. The evaporation products of LSA-SAF and MOD16 have again rather low values for the full range of catchments and do not add significant information (Figure 8e). At the same time, Budyko helps a lot, pointing also here at difficulties of the model to reproduce long-term behavior.

3.4. Added Value of Remote Sensing Data for Hydrological Signatures

Figure 11 summarizes all empirical cumulative distribution curves obtained from the combined significant KS statistics D^+ of all tested hydrological signatures for all combinations of products, relative to the unconstrained models. It can be observed that all models experience, on average, a shift toward higher values of the test statistic D^+ (i.e., to the right), when more products are included, also pointing toward improved model internal dynamics.

For FLEX and FLEXtopo (Figures 11a and 11b) the results suggest that a strong reduction in the search space can, on average, be achieved by including more remote sensing products compared to the unconstrained situation. This can be seen by the large shifts toward higher values between the envelopes of one product (gray) and more products (dark red colors). The final set of constraints, with all products included (red in Figure 11), is for FLEX close to containing the largest KS distances from the unconstrained situation but apparently not the set of constraints with the largest KS statistics D^+. Similarly, FLEXtopo also has large values for the test statistic for the final set of constraints, but this curve (red line) is, also here, not the curve with the biggest difference with the unconstrained case. This indicates that at least one of the products is not adding more value and actually reduces the models ability to reproduce the set of hydrological signatures. Inspection of the individual curves for FLEX shows that the curve, for nine products included, with the largest KS statistics D^+ is the curve without NDII. In Figure 7, this can also be observed in some cases, but the negative influence becomes much more apparent when evaluating a set of signatures. The curve with the lowest values for the KS statistic D^+ for nine products is however the curve without GRACE, pointing also at the importance of GRACE for reproducing the signatures. The additional gains by including GRACE are also in agreement with the findings of Rakovec, Kumar, Attinger, et al. (2016), who found that evaporation estimates largely improved.

For HYMOD (Figure 11c), the solutions obtained from the highest number of remote sensing products used to constrain ($n = 9$) are not the curves with the highest values for the KS statistics relative to the unconstrained solutions. The curve most toward high KS statistics D^+ when one data source is included (gray) is in this case Budyko, also in agreement with the previous findings for the objective function values. Most of the other curves for the individual products plot around a values of 0, indicating that it does not significantly influence the model results if this data source is included or not. After including Budyko in the combinations, these curves start to shift to the right, eventually leading to curves that end up with strong improvements due to including the Budyko framework in the set of constraints.

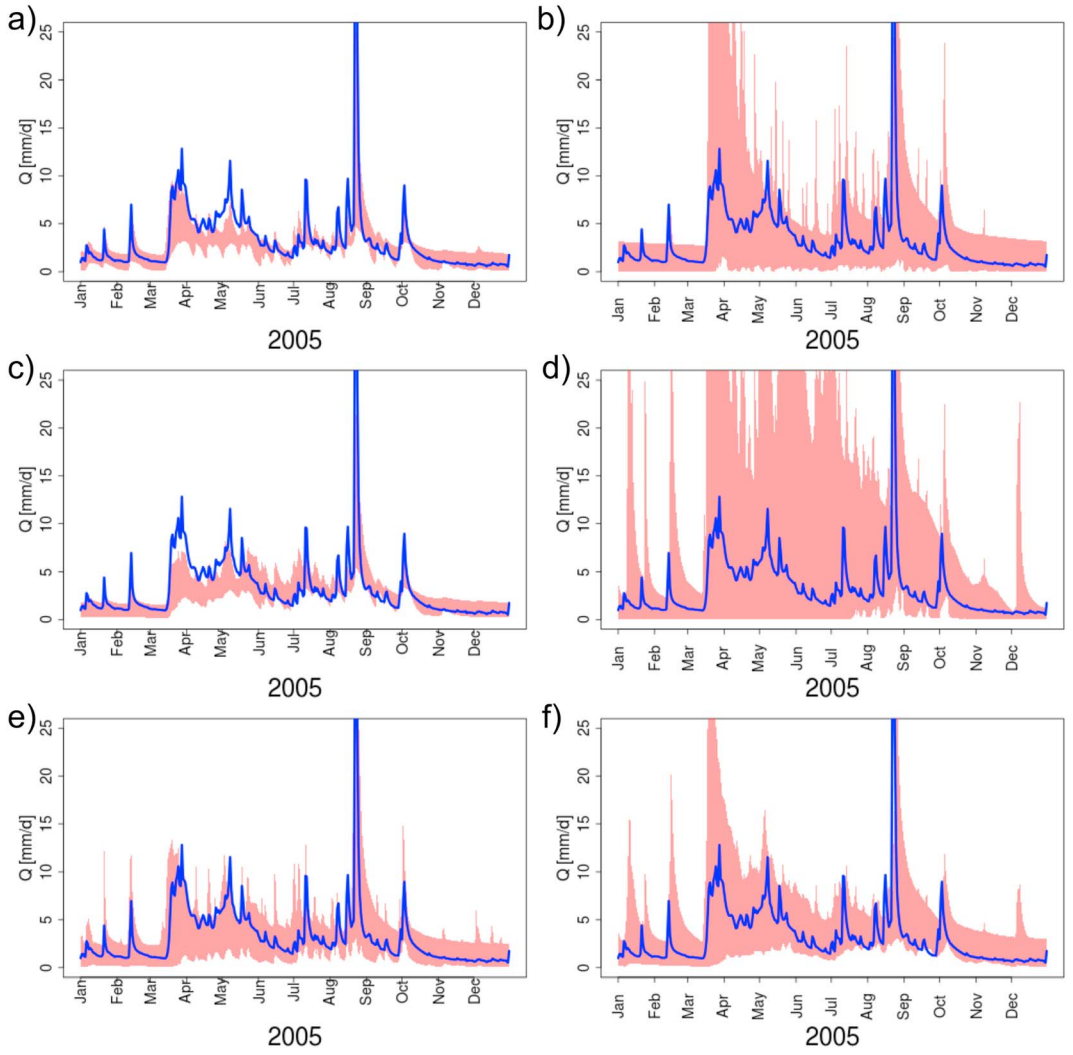

Figure 10. Feasible flow ranges for a selected time period of the Vils catchment (catchment 4) obtained with combinations of remote sensing products of (a) *with* MOD10 and FLEX, (b) *with* MOD10 and HYMOD, (c) *without* MOD10 and FLEX, (d) *without* MOD10 and HYMOD, (e) FLEX in the benchmark situation, and (f) HYMOD in the benchmark situation. Observed discharge is shown in blue, and the colored envelopes in Figures 10a–10d represent the feasible ranges with nine products used in deriving the posterior parameter distributions and, in Figures 10e and 10f, the benchmark situation.

The curves for HYPE (Figure 11d) are different compared to the curves of FLEX and FLEXtopo, as the envelopes of the curves resulting from the use of nine products are much narrower. Thus, the same number of products is much more restrictive for HYPE, leading to reduced uncertainty intervals. Even though the final set of constraints is not the curve with the highest values for the KS statistics, the curves are rather close to each other. Thus, excluding a certain product for the set used for constraining does not make a big difference, pointing at the combined strength of the other remaining nine products.

Also for TUW, the curves of nine products are relatively close to each other. In addition, the highest values for the KS statistics, relative to the unconstrained situation, are obtained when only a single product is used. Hence, when combinations are made, some constraints are relaxed and corrected by including other data sources, shifting the curves toward higher KS statistics D^+, other constraints become less effective, and shifting the curves toward lower KS statistics. In the end, the envelopes of the higher number of products are contained within the envelopes of the single products (gray), indicating that still all data sources helped to improve the representation of catchment signatures compared to the unconstrained situation. This points at the combined strength of the products, correcting too restrictive constraints and improving the signature representation together.

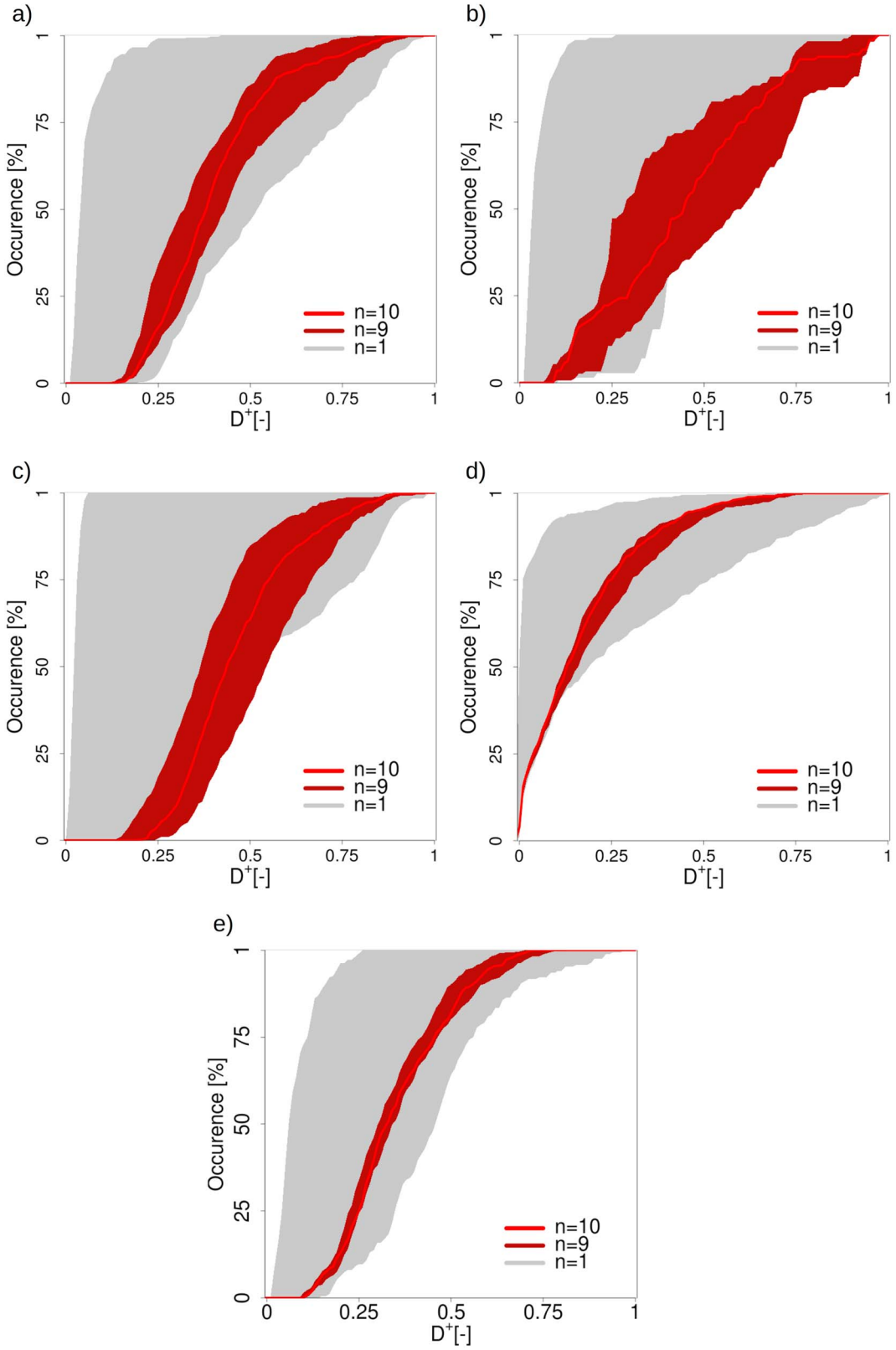

Figure 11. Envelopes of all empirical cumulative distribution curves for all significant KS statistics D^+ for all catchment signatures (each catchment, each signature) for the different combinations of constraints (i.e., remote sensing products) compared to the unconstrained model: (a) FLEX, (b) FLEXtopo, (c) HYMOD, (d) HYPE, and (e) TUW. The different colors represent the different number of remote sensing products used to constraining the models.

3.5. Limitations and Outlook

The presented research focused on applying many combinations of additional data sources to derive new parameter ranges and to constrain the feasible parameter space for five different models, where other studies used multiple additional data sources to either evaluate one specific model for at most three products (e.g., Rakovec, Kumar, Mai, et al., 2016), or constrain models for one specific set of products (e.g., Kunnath-Poovakka et al., 2016; Lopez Lopez et al., 2017). However, a limitation of this study, due to computational constraints and the elevated volume of produced data (~3 Terabyte), is that the remotely sensed products as well as models were all applied in a lumped, catchment-averaged manner. In this way the remote sensing products are, naturally, not used up to their full potential, being distributed data sources. In addition, even though consistent for both models and products, it can be argued that the spatial average for an entire area and the way it was defined (e.g., arithmetic mean and harmonic mean) may be unrepresentative (e.g., Kumar, Samaniego, et al., 2013; Kumar, Livneh, et al., 2013). Hence, a follow up on this study would ideally put more emphasis on the effect of alternative spatial aggregation methods, specifically for the states/fluxes under consideration, as well as more distributed modeling approaches to more completely exploit the information content of the products. This was already shown to be promising by several authors (Demirel et al., 2018; Zink et al., 2018), and using such approaches with combinations of multiple products, correctly linked to (spatially distributed) model states and fluxes, seems therefore an important research line to pursue.

The results in this study strongly depend, just as in any other study, on the quality of the used data. The relative errors in the remote sensing products can be considerable, just as the errors in the used time series of river discharge. These errors can also differ per specific place, depending on the original data supplier or the climatological situation. Ideally, analyses as presented here would focus on many catchments with high-quality data, which was attempted here with the selection of 27 catchments.

The choice of models remains an additional, subjective choice, influencing the presented results. Especially as several models have a large parameter space, a large number of samples is needed to obtain robust results. The 100,000 samples used in this study were merely imposed by the technical possibilities, but this number would ideally even be higher. This study is however about reducing the parameter search space, also in order to minimize sampling efforts, and the results still show that with a relatively low number of samples, improvements (e.g., reductions in the parameter search space) can still be achieved.

The measure of fit between modeled and remotely sensed states/fluxes in this research was generally the squared correlation coefficient. This was done on purpose as it is a relatively weak measure, which ignores biases and only assumes a linear relation. Nevertheless, it can be argued that most of the added value of the products comes forward from an improved seasonality of the models. However, the seasonal signals from the different products will still contain a different information component. For example, the seasonal signal of snow is totally different from the seasonal signal of soil moisture depletion or groundwater storage. In other words, the seasonal signals of the different individual products and thus the associated variables are considerably distinct in timing (i.e., phase shift), and only if superimposed onto each other produce the overall hydrological response. Removing seasonality from the modeled and observed data can be considered as well (see also Figure S7 in the supporting information), but the relative influence of data and model errors will be increased at the same time, leading to a very restrictive measure of fit. In Figure S7 it can be clearly seen, for example, that removing seasonality leads to rather low performances for the soil moisture products. This is generally a known issue, as these products represent the soil moisture in the top centimeters, whereas the models generally contain a complete bucket of soil moisture. Therefore, most of the information that the remotely sensed products added when constraining the models is probably the seasonal signal, as this was not removed, but, as stated before, this signal remains specific for each product. Therefore, it can be argued that the seasonal signals of the models and products are, or should be, similar. In future studies, a more proper linkage between the model state and remotely sensed variable is needed, instead of the linear correlation used here. As also the quality of the remote sensing products is increasing, the added value will also go beyond adding seasonality. Naturally, distributed model approaches may help in this aspect as well, considering the distributed nature of the products. Concluding, the choice of the used objective functions, either deseasonalized or not, remains a subjective choice, with a large influence on the results presented here.

4. Conclusions

Twenty-seven catchments across Europe were constrained with combinations of nine remotely sensed products and an analytical framework (Budyko). New posterior parameter distributions for five different conceptual hydrological models were derived based on a likelihood weighting procedure, which was specific for each parameter depending on the relevant data sources for that parameter. In this way, all 1,023 possible combinations of these 10 data sources could be used to derive new parameter ranges.

It was found that strong improvements were obtained when combinations included in particular AMSR-E and ASCAT soil moisture data. Surprisingly, considering the relatively small size of the study catchments, also GRACE added considerable value to meaningfully constrain the tested models. In addition, in snow-dominated catchments the MODIS snow products were shown to be helpful for some of these models. The evaporation products of LSA-SAF and MOD16 were to a lesser extent important for deriving adequate and meaningful posterior parameter distributions.

A set of 27 hydrological signatures was evaluated for each study catchment, and the improvement for reproducing these signatures using only remote sensing data for constraining a model compared to unconstrained models was analyzed by the KS statistic D^+ between the constrained and unconstrained models. This showed that all models benefitted from using a combination of remote sensing data for reproducing catchment signatures.

This study illustrates that using combinations of multiple data sources is in most cases valuable to derive reasonably narrow and meaningful posterior parameter distributions. It was shown that the highest gains are, on average, obtained when the soil moisture products AMSR-E and ASCAT as well as the total water storage anomaly of GRACE are included. Using multiple products simultaneously also corrects errors of a single product, and including four to five different products is often sufficient to reduce the parameter search space and gets closer to solutions obtained by a range of benchmarking solutions. In addition, hydrological signatures were better represented when multiple data sources were used, indicating improved model internal dynamics. In conclusion, adding multiple data sources in parameter selection procedures in an indirect, parameter specific way is a promising way forward in predicting ungauged catchments.

Acknowledgments

We would like to acknowledge the European Commission FP7 funded research project *Sharing Water-related Information to Tackle Changes in the Hydrosphere - for Operational Needs* (SWITCH-ON, grant agreement 603587), as this study was conducted within the context of SWITCH-ON as an example of scientific potential when using open data for collaborative research in hydrology. We acknowledge the data providers of The Global Runoff Data Centre, D-56002 Koblenz, Germany for providing discharge data (www.bafg.de), as well as the Hydrographic Service of Austria, the Hydrographic Service of the Autonomous Province of Bolzano, Regional Agency for the Protection of the Environment - Piedmont Region and Regional Hydrologic Service - Tuscany Region, the ECMWF for the ERA-Interim data (http://apps.ecmwf.int/datasets/data/interim-full-daily/), and the providers of the MSWEP for the precipitation data (http://www.gloh2o.org/). We acknowledge the Vrije Universiteit Amsterdam and NASA GSFC for the data of AMSR-E LPRM (https://gcmd.nasa.gov/KeywordSearch/Metadata.do?Portal=GCMD&KeywordPath=Parameters%7CLAND+SURFACE%7CSOILS%7CSOIL+MOISTURE%2FWATER+CONTENT&EntryId=GES_DISC_LPRM_AMSRE_A_SOILM3_V002), Copernicus Land Products and TU Vienna for the data of ASCAT-SWI (http://land.copernicus.eu/global/), the ESA for the SMOS data set (https://earth.esa.int/web/guest/missions/esa-operational-eo-missions/smos), and the NASA National Snow and Ice Data Center Distributed Active Archive Center for the MODIS Terra and Aqua data sets of MOD10 and MYD10 (https://modis.gsfc.nasa.gov/data/dataprod/mod10.php). The MOD09 data product was retrieved from the online Data Pool, courtesy of the NASA Land Processes Distributed Active Archive Center (LP DAAC), USGS/Earth Resources Observation and Science (EROS) Center, Sioux Falls, South Dakota (https://lpdaac.usgs.gov/data_access/data_pool) and Numerical Terradynamic Simulation Group for the MOD16 data (http://www.ntsg.umt.edu/project/modis/mod16.php). The LSA-SAF evaporation was provided by the EUMETSAT Satellite Application Facility on Land Surface Analysis (http://landsaf.ipma.pt). GRACE data are available at http://grace.jpl.nasa.gov, supported by the NASA MEaSUREs Program. We would like to thank the reviewers, Editor, and Associate Editor for their reviews and constructive feedback.

References

AghaKouchak, A., Farahmand, A., Melton, F. S., Teixeira, J., Anderson, M. C., Wardlow, B. D., & Hain, C. R. (2015). Remote sensing of drought: Progress, challenges and opportunities. *Reviews of Geophysics*, *53*, 452–480. https://doi.org/10.1002/2014RG000456

Allen, R. G., L. S. Pereira, D. Raes, and M. Smith (1998). Crop evapotranspiration-guidelines for computing crop water requirements. FAO irrigation and drainage paper 56, Rome: FAO.

Almeida, S., Le Vine, N., McIntyre, N., Wagener, T., & Buytaert, W. (2016). Accounting for dependencies in regionalized signatures for predictions in ungauged catchments. *Hydrology and Earth System Sciences*, *20*(2), 887–901. https://doi.org/10.5194/hess-20-887-2016

Ambroise, B., Beven, K., & Freer, J. (1996). Toward a generalization of the TOPMODEL concepts: Topographic indices of hydrological similarity. *Water Resources Research*, *32*(7), 2135–2145. https://doi.org/10.1029/95WR03716

Andréassian, V., Le Moine, N., Perrin, C., Ramos, M.-H., Oudin, L., Mathevet, T., et al. (2012). All that glitters is not gold: The case of calibrating hydrological models. *Hydrological Processes*, *26*(14), 2206–2210. https://doi.org/10.1002/hyp.9264

Bárdossy, A. (2007). Calibration of hydrological model parameters for ungauged catchments. *Hydrology and Earth System Sciences*, *11*(2), 703–710. https://doi.org/10.5194/hess-11-703-2007

Bartalis, Z., Wagner, W., Naeimi, V., Hasenauer, S., Scipal, K., Bonekamp, H., et al. (2007). Initial soil moisture retrievals from the METOP-A Advanced Scatterometer (ASCAT). *Geophysical Research Letters*, *34*, L20401. https://doi.org/10.1029/2007GL031088

Beck, H. E., van Dijk, A. I. J. M., Levizzani, V., Schellekens, J., Miralles, D. G., Martens, B., & de Roo, A. (2017). MSWEP: 3-hourly 0.25° global gridded precipitation (1979–2015) by merging gauge, satellite, and reanalysis data. *Hydrology and Earth System Sciences*, *21*(1), 589–615.

Bergström, S. (1992). *The HBV model: Its structure and applications*. Stockholm: Swedish Meteorological and Hydrological Institute.

Beven, K. (2006). A manifesto for the equifinality thesis. *Journal of Hydrology*, *320*(1-2), 18–36. https://doi.org/10.1016/j.jhydrol.2005.07.007

Beven, K., & Freer, J. (2001). Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology. *Journal of Hydrology*, *249*(1-4), 11–29. https://doi.org/10.1016/S0022-1694(01)00421-8

Blöschl, G., Sivapalan, M., Wagener, T., Viglione, A., & Savenije, H. H. G. (2013). *Runoff prediction in ungauged basins: Synthesis across processes, places and scales*. Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9781139235761

de Boer-Euser, T., Bouaziz, L., de Niel, J., Brauer, C., Dewals, B., Drogue, G., et al. (2017). Looking beyond general metrics for model comparison—Lessons from an international model intercomparison study. *Hydrology and Earth System Sciences*, *21*(1), 423–440. https://doi.org/10.5194/hess-21-423-2017

de Boer-Euser, T., McMillan, H. K., Hrachowitz, M., Winsemius, H. C., & Savenije, H. H. G. (2016). Influence of soil and climate on root zone storage capacity. *Water Resources Research*, *52*, 2009–2024. https://doi.org/10.1002/2015WR018115

Boyle, D. P. (2001). *Multicriteria calibration of hydrologic models*. Tucson: Dep. of Hydrol. and Water Resour., Univ. of Ariz.

Brocca, L., Melone, F., Moramarco, T., Wagner, W., Naeimi, V., Bartalis, Z., & Hasenauer, S. (2010). Improving runoff prediction through the assimilation of the ASCAT soil moisture product. *Hydrology and Earth System Sciences*, *14*(10), 1881–1893. https://doi.org/10.5194/hess-14-1881-2010

Brown, M. E., Escobar, V., Moran, S., Entekhabi, D., O'Neill, P. E., Njoku, E. G., et al. (2013). NASA's Soil Moisture Active Passive (SMAP) mission and opportunities for applications users. *Bulletin of the American Meteorological Society*, *94*(8), 1125–1128. https://doi.org/10.1175/BAMS-D-11-00049.1

Budyko, M. I. (1974). *Climate and life*. San Diego, CA: Academic Press.

Bulygina, N., McIntyre, N., & Wheater, H. (2009). Conditioning rainfall-runoff model parameters for ungauged catchments and land management impacts analysis. *Hydrology and Earth System Sciences*, *13*(6), 893–904. https://doi.org/10.5194/hess-13-893-2009

Castiglioni, S., Castellarin, A., Montanari, A., Skøien, J. O., Laaha, G., & Blöschl, G. (2011). Smooth regional estimation of low-flow indices: Physiographical space based interpolation and top-kriging. *Hydrology and Earth System Sciences*, *15*(3), 715–727. https://doi.org/10.5194/hess-15-715-2011

Castiglioni, S., Lombardi, L., Toth, E., Castellarin, A., & Montanari, A. (2010). Calibration of rainfall-runoff models in ungauged basins: A regional maximum likelihood approach. *Advances in Water Resources*, *33*(10), 1235–1242. https://doi.org/10.1016/j.advwatres.2010.04.009

Ceola, S., Arheimer, B., Baratti, E., Blöschl, G., Capell, R., Castellarin, A., et al. (2015). Virtual laboratories: New opportunities for collaborative water science. *Hydrology and Earth System Sciences*, *19*(4), 2101–2117. https://doi.org/10.5194/hess-19-2101-2015

Crow, W. T., & Ryu, D. (2009). A new data assimilation approach for improving runoff prediction using remotely-sensed soil moisture retrievals. *Hydrology and Earth System Sciences*, *13*(1), 1–16. https://doi.org/10.5194/hess-13-1-2009

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., et al. (2011). The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, *137*(656), 553–597. https://doi.org/10.1002/qj.828

Demaria, E. M., Nijssen, B., & Wagener, T. (2007). Monte Carlo sensitivity analysis of land surface parameters using the Variable Infiltration Capacity model. *Journal of Geophysical Research*, *112*, D11113. https://doi.org/10.1029/2006JD007534

Demirel, M. C., Mai, J., Mendiguren, G., Koch, J., Samaniego, L., & Stisen, S. (2018). Combining satellite data and appropriate objective functions for improved spatial pattern performance of a distributed hydrologic model. *Hydrology and Earth System Sciences*, *22*(2), 1299–1315. https://doi.org/10.5194/hess-22-1299-2018

Donnelly, C., Dahne, J., Lindström, G., Rosberg, J., Strömqvist, J., Pers, C., et al. (2009). An evaluation of multi-basin hydrological modelling for predictions in ungauged basins. In K. Yilmaz, et al. (Eds.), *New approaches to hydrological prediction in data sparse regions*, (pp. 112–120). Wallingford: IAHS Press.

Duan, Q., Sorooshian, S., & Gupta, V. (1992). Effective and efficient global optimization for conceptual rainfall-runoff models. *Water Resources Research*, *28*(4), 1015–1031. https://doi.org/10.1029/91WR02985

Euser, T., Winsemius, H. C., Hrachowitz, M., Fenicia, F., Uhlenbrook, S., & Savenije, H. H. G. (2013). A framework to assess the realism of model structures using hydrological signatures. *Hydrology and Earth System Sciences*, *17*(5), 1893–1912. https://doi.org/10.5194/hess-17-1893-2013

Fekete, B. M., and C. J. Vörösmarty (2002). The current status of global river discharge monitoring and potential new technologies complementing traditional discharge measurements. Paper presented at Predictions in Ungauged Basins: PUB kick-off (Proceedings of the PUB kick-off meeting held in Brasilia, 20–22 November 2002). IAHS Publication.

Fenicia, F., Savenije, H. H. G., Matgen, P., & Pfister, L. (2008). Understanding catchment behavior through stepwise model concept improvement. *Water Resources Research*, *44*, W01402. https://doi.org/10.1029/2006WR005563

Fovet, O., Ruiz, L., Hrachowitz, M., Faucheux, M., & Gascuel-Odoux, C. (2015). Hydrological hysteresis and its value for assessing process consistency in catchment conceptual models. *Hydrology and Earth System Sciences*, *19*(1), 105–123.

Freer, J., Beven, K., & Ambroise, B. (1996). Bayesian estimation of uncertainty in runoff prediction and the value of data: An application of the GLUE approach. *Water Resources Research*, *32*(7), 2161–2173. https://doi.org/10.1029/95WR03723

Freer, J. E., McMillan, H., McDonnell, J. J., & Beven, K. J. (2004). Constraining dynamic TOPMODEL responses for imprecise water table information using fuzzy rule based performance measures. *Journal of Hydrology*, *291*(3-4), 254–277. https://doi.org/10.1016/j.jhydrol.2003.12.037

Gao, H., Hrachowitz, M., Fenicia, F., Gharari, S., & Savenije, H. H. G. (2014). Testing the realism of a topography-driven model (FLEX-Topo) in the nested catchments of the upper Heihe, China. *Hydrology and Earth System Sciences*, *18*(5), 1895–1915. https://doi.org/10.5194/hess-18-1895-2014

Gao, H., Hrachowitz, M., Schymanski, S. J., Fenicia, F., Sriwongsitanon, N., & Savenije, H. H. G. (2014). Climate controls how ecosystems size the root zone storage capacity at catchment scale. *Geophysical Research Letters*, *41*, 7916–7923. https://doi.org/10.1002/2014GL061668

Gentine, P., D'Odorico, P., Lintner, B. R., Sivandran, G., & Salvucci, G. (2012). Interdependence of climate, soil, and vegetation as constrained by the Budyko curve. *Geophysical Research Letters*, *39*, L19404. https://doi.org/10.1029/2012GL053492

Gerrits, A. M. J., Savenije, H. H. G., Veling, E. J. M., & Pfister, L. (2009). Analytical derivation of the Budyko curve based on rainfall characteristics and a simple evaporation model. *Water Resources Research*, *45*, W04403. https://doi.org/10.1029/2008WR007308

Gharari, S., Hrachowitz, M., Fenicia, F., Gao, H., & Savenije, H. H. G. (2014). Using expert knowledge to increase realism in environmental system models can dramatically reduce the need for calibration. *Hydrology and Earth System Sciences*, *18*(12), 4839–4859. https://doi.org/10.5194/hess-18-4839-2014

Ghilain, N., Arboleda, A., & Gellens-Meulenberghs, F. (2011). Evapotranspiration modelling at large scale using near-real time MSG SEVIRI derived data. *Hydrology and Earth System Sciences*, *15*(3), 771–786. https://doi.org/10.5194/hess-15-771-2011

Githui, F., Thayalakumaran, T., & Selle, B. (2015). Estimating irrigation inputs for distributed hydrological modelling: A case study from an irrigated catchment in southeast Australia. *Hydrological Processes*, *30*(12), 1824–1835.

Götzinger, J., & Bárdossy, A. (2007). Comparison of four regionalisation methods for a distributed hydrological model. *Journal of Hydrology*, *333*(2–4), 374–384. https://doi.org/10.1016/j.jhydrol.2006.09.008

Greenstone, R., & King, M. D. (1999). *EOS science plan: Executive summary*. Greenbelt, MD: NASA GSFC.

Hall, D. K., Salomonson, V. V., & Riggs, G. A. (2006a). *MODIS/Terra snow cover daily L3 global 500m grid*. Boulder, CO: National Snow and Ice Data Center (Version 5).

Hall, D. K., Salomonson, V. V., & Riggs, G. A. (2006b). *MODIS/Aqua snow cover daily L3 global 500m grid*. Boulder, CO: National Snow and Ice Data Center (Version 5).

Hannah, D. M., Demuth, S., van Lanen, H. A. J., Looser, U., Prudhomme, C., Rees, G., et al. (2011). Large-scale river flow archives: Importance, current status and future needs. *Hydrological Processes*, *25*(7), 1191–1200. https://doi.org/10.1002/hyp.7794

Hornberger, G. M., & Spear, R. C. (1980). Eutrophication in peel inlet—I. The problem-defining behavior and a mathematical model for the phosphorus scenario. *Water Research*, *14*(1), 29–42.

Hrachowitz, M., & Clark, M. P. (2017). HESS opinions: The complementary merits of competing modelling philosophies in hydrology. *Hydrology and Earth System Sciences*, *21*(8), 3953–3973. https://doi.org/10.5194/hess-21-3953-2017

Hrachowitz, M., Fovet, O., Ruiz, L., Euser, T., Gharari, S., Nijzink, R., et al. (2014). Process consistency in models: The importance of system signatures, expert knowledge, and process complexity. *Water Resources Research*, *50*, 7445–7469. https://doi.org/10.1002/2014WR015484
Hrachowitz, M., Savenije, H. H. G., Blöschl, G., McDonnell, J. J., Sivapalan, M., Pomeroy, J. W., et al. (2013). A decade of predictions in ungauged basins (PUB)—A review. *Hydrological Sciences Journal*, *58*(6), 1198–1255. https://doi.org/10.1080/02626667.2013.803183
Hundecha, Y., Arheimer, B., Donnelly, C., & Pechlivanidis, I. (2016). A regional parameter estimation scheme for a pan-European multi-basin model. *Journal of Hydrology: Regional Studies*, *6*, 90–111.
Hundecha, Y., & Bárdossy, A. (2004). Modeling of the effect of land use changes on the runoff generation of a river basin through parameter regionalization of a watershed model. *Journal of Hydrology*, *292*(1–4), 281–295. https://doi.org/10.1016/j.jhydrol.2004.01.002
Immerzeel, W. W., & Droogers, P. (2008). Calibration of a distributed hydrological model based on satellite evapotranspiration. *Journal of Hydrology*, *349*(3-4), 411–424. https://doi.org/10.1016/j.jhydrol.2007.11.017
Jothityangkoon, C., Sivapalan, M., & Farmer, D. L. (2001). Process controls of water balance variability in a large semi-arid catchment: Downward approach to hydrological model development. *Journal of Hydrology*, *254*(1-4), 174–198. https://doi.org/10.1016/S0022-1694(01)00496-6
Kelleher, C., McGlynn, B., & Wagener, T. (2017). Characterizing and reducing equifinality by constraining a distributed catchment model with regional signatures, local observations, and process understanding. *Hydrology and Earth System Sciences*, *21*(7), 3325–3352. https://doi.org/10.5194/hess-21-3325-2017
Kerr, Y. H., Waldteufel, P., Richaume, P., Wigneron, J. P., Ferrazzoli, P., Mahmoodi, A., et al. (2012). The SMOS soil moisture retrieval algorithm. *IEEE Transactions on Geoscience and Remote Sensing*, *50*(5), 1384–1403. https://doi.org/10.1109/TGRS.2012.2184548
Kirchner, J. W. (2006). Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology. *Water Resources Research*, *42*, W03S04. https://doi.org/10.1029/2005WR004362
Kuentz, A., Arheimer, B., Hundecha, Y., & Wagener, T. (2017). Understanding hydrologic variability across Europe through catchment classification. *Hydrology and Earth System Sciences*, *21*(6), 2863–2879. https://doi.org/10.5194/hess-21-2863-2017
Kumar, R., Livneh, B., & Samaniego, L. (2013). Toward computationally efficient large-scale hydrologic predictions with a multiscale regionalization scheme. *Water Resources Research*, *49*, 5700–5714. https://doi.org/10.1002/wrcr.20431
Kumar, R., Samaniego, L., & Attinger, S. (2013). Implications of distributed hydrologic model parameterization on water fluxes at multiple scales and locations. *Water Resources Research*, *49*, 360–379. https://doi.org/10.1029/2012WR012195
Kunnath-Poovakka, A., Ryu, D., Renzullo, L. J., & George, B. (2016). The efficacy of calibrating hydrologic model using remotely sensed evapotranspiration and soil moisture for streamflow prediction. *Journal of Hydrology*, *535*, 509–524. https://doi.org/10.1016/j.jhydrol.2016.02.018
Landerer, F. W., & Swenson, S. C. (2012). Accuracy of scaled GRACE terrestrial water storage estimates. *Water Resources Research*, *48*, W04531. https://doi.org/10.1029/2011WR011453
Lettenmaier, D. P., Alsdorf, D., Dozier, J., Huffman, G. J., Pan, M., & Wood, E. F. (2015). Inroads of remote sensing into hydrologic science during the WRR era. *Water Resources Research*, *51*, 7309–7342. https://doi.org/10.1002/2015WR017616
Li, H.-Y., Sivapalan, M., Tian, F., & Harman, C. (2014). Functional approach to exploring climatic and landscape controls of runoff generation: 1. Behavioral constraints on runoff volume. *Water Resources Research*, *50*, 9300–9322. https://doi.org/10.1002/2014WR016307
Lindström, G., Pers, C., Rosberg, J., Strömqvist, J., & Arheimer, B. (2010). Development and testing of the HYPE (Hydrological Predictions for the Environment) water quality model for different spatial scales. *Hydrology Research*, *41*(3–4), 295–319. https://doi.org/10.2166/nh.2010.007
Liu, Y., & Gupta, H. V. (2007). Uncertainty in hydrologic modeling: Toward an integrated data assimilation framework. *Water Resources Research*, *43*, W07401. https://doi.org/10.1029/2006WR005756
Liu, Y., Weerts, A. H., Clark, M., Hendricks Franssen, H. J., Kumar, S., Moradkhani, H., et al. (2012). Advancing data assimilation in operational hydrologic forecasting: Progresses, challenges, and emerging opportunities. *Hydrology and Earth System Sciences*, *16*(10), 3863–3887. https://doi.org/10.5194/hess-16-3863-2012
Lo, M.-H., Famiglietti, J. S., Yeh, P. J. F., & Syed, T. H. (2010). Improving parameter estimation and water table depth simulation in a land surface model using GRACE water storage and estimated base flow data. *Water Resources Research*, *46*, W05517. https://doi.org/10.1029/2009WR007855
Lopez Lopez, P., Sutanudjaja, E., Schellekens, J., Sterk, G., & Bierkens, M. (2017). Calibration of a large-scale hydrological model using satellite-based soil moisture and evapotranspiration products. *Hydrology and Earth System Sciences Discussions*, *21*, 3125–3144.
McIntyre, N. R., Wagener, T., Wheater, H. S., & Chapra, S. C. (2003). Risk-based modelling of surface water quality: A case study of the Charles River, Massachusetts. *Journal of Hydrology*, *274*(1-4), 225–247. https://doi.org/10.1016/S0022-1694(02)00417-1
Merz, R., & Blöschl, G. (2004). Regionalisation of catchment model parameters. *Journal of Hydrology*, *287*(1-4), 95–123. https://doi.org/10.1016/j.jhydrol.2003.09.028
Montanari, A., & Toth, E. (2007). Calibration of hydrological models in the spectral domain: An opportunity for scarcely gauged basins? *Water Resources Research*, *43*, W05434. https://doi.org/10.1029/2006WR005184
Mu, Q., Zhao, M., & Running, S. W. (2011). Improvements to a MODIS global terrestrial evapotranspiration algorithm. *Remote Sensing of Environment*, *115*(8), 1781–1800. https://doi.org/10.1016/j.rse.2011.02.019
Mulder, G., Olsthoorn, T. N., Al-Manmi, D. A. M. A., Schrama, E. J. O., & Smidt, E. H. (2015). Identifying water mass depletion in northern Iraq observed by GRACE. *Hydrology and Earth System Sciences*, *19*(3), 1487–1500. https://doi.org/10.5194/hess-19-1487-2015
Nijzink, R., et al. (2016). The evolution of root-zone moisture capacities after deforestation: A step towards hydrological predictions under change? *Hydrology and Earth System Sciences*, *20*(12), 4775–4799.
Oudin, L., Andréassian, V., Perrin, C., & Anctil, F. (2004). Locating the sources of low-pass behavior within rainfall-runoff models. *Water Resources Research*, *40*, W11101. https://doi.org/10.1029/2004WR003291
Owe, M., de Jeu, R., & Holmes, T. (2008). Multisensor historical climatology of satellite-derived global land surface moisture. *Journal of Geophysical Research*, *113*, F01002. https://doi.org/10.1029/2007JF000769
Parajka, J., & Blöschl, G. (2008). The value of MODIS snow cover data in validating and calibrating conceptual hydrologic models. *Journal of Hydrology*, *358*(3-4), 240–258. https://doi.org/10.1016/j.jhydrol.2008.06.006
Parajka, J., Merz, R., & Blöschl, G. (2007). Uncertainty and multiple objective calibration in regional water balance modelling: Case study in 320 Austrian catchments. *Hydrological Processes*, *21*(4), 435–446. https://doi.org/10.1002/hyp.6253
Parajka, J., Naeimi, V., Blöschl, G., & Komma, J. (2009). Matching ERS scatterometer based soil moisture patterns with simulations of a conceptual dual layer hydrologic model over Austria. *Hydrology and Earth System Sciences*, *13*(2), 259–271. https://doi.org/10.5194/hess-13-259-2009
Pechlivanidis, I. G., & Arheimer, B. (2015). Large-scale hydrological modelling by using modified PUB recommendations: The India-HYPE case. *Hydrology and Earth System Sciences*, *19*(11), 4559–4579. https://doi.org/10.5194/hess-19-4559-2015

Rakovec, O., Kumar, R., Attinger, S., & Samaniego, L. (2016). Improving the realism of hydrologic model functioning through multivariate parameter estimation. *Water Resources Research*, *52*, 7779–7792. https://doi.org/10.1002/2016WR019430

Rakovec, O., Kumar, R., Mai, J., Cuntz, M., Thober, S., Zink, M., et al. (2016). Multiscale and multivariate evaluation of water fluxes and states over European River basins. *Journal of Hydrometeorology*, *17*(1), 287–307. https://doi.org/10.1175/JHM-D-15-0054.1

Reichle, R. H. (2008). Data assimilation methods in the Earth sciences. *Advances in Water Resources*, *31*(11), 1411–1418. https://doi.org/10.1016/j.advwatres.2008.01.001

Samaniego, L., Kumar, R., & Attinger, S. (2010). Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale. *Water Resources Research*, *46*, W05523. https://doi.org/10.1029/2008WR007327

Savenije, H. H. G. (2010). HESS opinions "topography driven conceptual modelling (FLEX-topo)". *Hydrology and Earth System Sciences*, *14*(12), 2681–2692. https://doi.org/10.5194/hess-14-2681-2010

Seibert, J., & McDonnell, J. J. (2002). On the dialog between experimentalist and modeler in catchment hydrology: Use of soft data for multicriteria model calibration. *Water Resources Research*, *38*(11), 1241. https://doi.org/10.1029/2001WR000978

Shamir, E., Imam, B., Morin, E., Gupta, H. V., & Sorooshian, S. (2005). The role of hydrograph indices in parameter estimation of rainfall–runoff models. *Hydrological Processes*, *19*(11), 2187–2207. https://doi.org/10.1002/hyp.5676

Sieber, A., & Uhlenbrook, S. (2005). Sensitivity analyses of a distributed catchment model to verify the model structure. *Journal of Hydrology*, *310*(1-4), 216–235. https://doi.org/10.1016/j.jhydrol.2005.01.004

Sivapalan, M. (2003). Prediction in ungauged basins: A grand challenge for theoretical hydrology. *Hydrological Processes*, *17*(15), 3163–3170. https://doi.org/10.1002/hyp.5155

Smith, T., Hayes, K., Marshall, L., McGlynn, B., & Jencso, K. (2016). Diagnostic calibration and cross-catchment transferability of a simple process-consistent hydrologic model. *Hydrological Processes*, *30*(26), 5027–5038. https://doi.org/10.1002/hyp.10955

Spaaks, J. H., & Bouten, W. (2013). Resolving structural errors in a spatially distributed hydrologic model using ensemble Kalman filter state updates. *Hydrology and Earth System Sciences*, *17*(9), 3455–3472. https://doi.org/10.5194/hess-17-3455-2013

Sriwongsitanon, N., Gao, H., Savenije, H. H. G., Maekan, E., Saengsawang, S., & Thianpopirug, S. (2016). Comparing the Normalized Difference Infrared Index (NDII) with root zone storage in a lumped conceptual model. *Hydrology and Earth System Sciences*, *20*(8), 3361–3377. https://doi.org/10.5194/hess-20-3361-2016

Stisen, S., Jensen, K. H., Sandholt, I., & Grimes, D. I. F. (2008). A remote sensing driven distributed hydrological model of the Senegal River basin. *Journal of Hydrology*, *354*(1-4), 131–148. https://doi.org/10.1016/j.jhydrol.2008.03.006

Sun, W., Ishidaira, H., Bastola, S., & Yu, J. (2015). Estimating daily time series of streamflow using hydrological model calibrated based on satellite observations of river water surface width: Toward real world applications. *Environmental Research*, *139*, 36–45. https://doi.org/10.1016/j.envres.2015.01.002

Sutanudjaja, E. H., van Beek, L. P. H., de Jong, S. M., van Geer, F. C., & Bierkens, M. F. P. (2014). Calibrating a large-extent high-resolution coupled groundwater-land surface model using soil moisture and discharge data. *Water Resources Research*, *50*, 687–705. https://doi.org/10.1002/2013WR013807

Tangdamrongsub, N., Steele-Dunne, S. C., Gunter, B. C., Ditmar, P. G., & Weerts, A. H. (2015). Data assimilation of GRACE terrestrial water storage estimates into a regional hydrological model of the Rhine River basin. *Hydrology and Earth System Sciences*, *19*(4), 2079–2100. https://doi.org/10.5194/hess-19-2079-2015

Tapley, B. D., Bettadpur, S., Watkins, M., & Reigber, C. (2004). The gravity recovery and climate experiment: Mission overview and early results. *Geophysical Research Letters*, *31*, L09607. https://doi.org/10.1029/2004GL019920

Tian, S., Tregoning, P., Renzullo, L. J., van Dijk, A. I. J. M., Walker, J. P., Pauwels, V. R. N., & Allgeyer, S. (2017). Improved water balance component estimates through joint assimilation of GRACE water storage and SMOS soil moisture retrievals. *Water Resources Research*, *53*, 1820–1840. https://doi.org/10.1002/2016WR019641

Tolson, B. A., & Shoemaker, C. A. (2007). Dynamically dimensioned search algorithm for computationally efficient watershed model calibration. *Water Resources Research*, *43*, W01413. https://doi.org/10.1029/2005WR004723

Tourian, M. J., Schwatke, C., & Sneeuw, N. (2017). River discharge estimation at daily resolution from satellite altimetry over an entire river basin. *Journal of Hydrology*, *546*, 230–247. https://doi.org/10.1016/j.jhydrol.2017.01.009

Vereecken, H., Huisman, J. A., Bogena, H., Vanderborght, J., Vrugt, J. A., & Hopmans, J. W. (2008). On the value of soil moisture measurements in vadose zone hydrology: A review. *Water Resources Research*, *44*, W00D06. https://doi.org/10.1029/2008WR006829

Verstraeten, W., Veroustraete, F., & Feyen, J. (2008). Assessment of evapotranspiration and soil moisture content across different scales of observation. *Sensors*, *8*(1), 70–117. https://doi.org/10.3390/s8010070

Wagener, T., Boyle, D. P., Lees, M. J., Wheater, H. S., Gupta, H. V., & Sorooshian, S. (2001). A framework for development and application of hydrological models. *Hydrology and Earth System Sciences*, *5*(1), 13–26. https://doi.org/10.5194/hess-5-13-2001

Wagener, T., & Kollat, J. (2007). Numerical and visual evaluation of hydrological and environmental models using the Monte Carlo analysis toolbox. *Environmental Modelling & Software*, *22*(7), 1021–1033. https://doi.org/10.1016/j.envsoft.2006.06.017

Wagener, T., & Montanari, A. (2011). Convergence of approaches toward reducing uncertainty in predictions in ungauged basins. *Water Resources Research*, *47*, W06301. https://doi.org/10.1029/2010WR009469

Wagener, T., & Wheater, H. S. (2006). Parameter estimation and regionalization for continuous rainfall-runoff models including uncertainty. *Journal of Hydrology*, *320*(1-2), 132–154. https://doi.org/10.1016/j.jhydrol.2005.07.015

Wagner, W., Lemoine, G., & Rott, H. (1999). A method for estimating soil moisture from ERS scatterometer and soil data. *Remote Sensing of Environment*, *70*(2), 191–207. https://doi.org/10.1016/S0034-4257(99)00036-X

Wanders, N., Bierkens, M. F. P., de Jong, S. M., de Roo, A., & Karssenberg, D. (2014). The benefits of using remotely sensed soil moisture in parameter identification of large-scale hydrological models. *Water Resources Research*, *50*, 6874–6891. https://doi.org/10.1002/2013WR014639

Werth, S., Güntner, A., Petrovic, S., & Schmidt, R. (2009). Integration of GRACE mass variations into a global hydrological model. *Earth and Planetary Science Letters*, *277*(1–2), 166–173. https://doi.org/10.1016/j.epsl.2008.10.021

Westerberg, I. K., Guerrero, J. L., Younger, P. M., Beven, K. J., Seibert, J., Halldin, S., et al. (2011). Calibration of hydrological models using flow-duration curves. *Hydrology and Earth System Sciences*, *15*(7), 2205–2227. https://doi.org/10.5194/hess-15-2205-2011

Winsemius, H. C., Schaefli, B., Montanari, A., & Savenije, H. H. G. (2009). On the calibration of hydrological models in ungauged basins: A framework for integrating hard and soft hydrological information. *Water Resources Research*, *45*, W12422. https://doi.org/10.1029/2009WR007706

Xu, X., Li, J., & Tolson, B. A. (2014). Progress in integrating remote sensing data and hydrologic modeling. *Progress in Physical Geography*, *38*(4), 464–498. https://doi.org/10.1177/0309133314536583

Yadav, M., Wagener, T., & Gupta, H. (2007). Regionalization of constraints on expected watershed response behavior for improved predictions in ungauged basins. *Advances in Water Resources*, *30*(8), 1756–1774. https://doi.org/10.1016/j.advwatres.2007.01.005

Yilmaz, K. K., Gupta, H. V., & Wagener, T. (2008). A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model. *Water Resources Research*, *44*, W09417. https://doi.org/10.1029/2007WR006716

Zhang, K., Kimball, J. S., & Running, S. W. (2016). A review of remote sensing based actual evapotranspiration estimation. *Wiley Interdisciplinary Reviews Water*, *3*(6), 834–853. https://doi.org/10.1002/wat2.1168

Zink, M., Mai, J., Cuntz, M., & Samaniego, L. (2018). Conditioning a hydrologic model using patterns of remotely sensed land surface temperature. *Water Resources Research*, *54*, 2976–2998. https://doi.org/10.1002/2017WR021346